KAERA
Research Forum

Special Issue:
Topics in Educational Measurement

Volume 1
Number 1
February 2014

Table of Contents

Forward..... 1

Editorial..... 2

DIF Analysis using a Mixture 3PL Model with a Covariate on the TIMSS 2007 Mathematics Test
Youn-Jeng Choi, Natalia Alexeev, and Allan S. Cohen..... 4

Application of Cognitive Diagnostic Model for Achievement Profile Analysis
HeeKyoung Kim..... 15

Linking with Constructed Response Items: A Hierarchical Model Approach with AP Data
YoungKoung Kim, Lawrence T. DeCarlo, and Rosemary Reshetar..... 26

The College Scholastic Ability Test in Korea: Introduction and Related Issues
Chanho Park 36

Analyzing Standard Setting Data Using a Generalizability Theory Framework
MinJeong Shin 47

Dealing with Measurement Error in Estimating a Cross-level Interaction: Nonlinear Multilevel Latent Variable Modeling Approach with a Metropolis-Hastings Robbins-Monro Algorithm
Ji Seung Yang and Li Cai..... 55

Call for Manuscripts..... 72

KAERA Research Forum, *1(1)*, 1

FORWARD

I am pleased to announce the launch of the *KAERA Research Forum,* an open access online scholarly forum, published by the Korean-American Educational Researchers' Association.

KAERA Research Forum is a research report series that discusses a variety of topics in educational research and disseminates high-quality examples of theoretical and empirical research studies to inform the larger community of educational researchers and practitioners. *KAERA Research Forum* intends to serve as the marketplace of ideas by publishing and disseminating information about the most up-to-date scholarly endeavors and experiments pursued by the members of KAERA community and beyond. Taking the form of research briefs online, it aims to facilitate speedy and efficient sharing of new ideas among KAERA members and educational researchers at large.

KAERA Research Forum gives an opportunity to both established scholars and emerging researchers including graduate students. For established scholars, this is a place where they can share their newest, intriguing ideas still in progress and unfolding. For junior scholars and graduate students, this forum will serve as one of first outlets to share and disseminate their scholarly work. *KAERA Research Forum* publishes four issues per year. Each issue presents a special theme and includes studies pertaining to a specific field of educational research or a particular topic or methodology. Editors for each special issue are selected and invited considering the scholarly expertise and leadership capacity of the individual scholar.

Last and most important, I would like to highlight that the launch of the *KAERA Research Forum* could not have been possible without the unwavering support of the KAERA Board of Directors, KAERA Executive Group, and several KAERA scholars who eagerly stepped up to serve as the guest editors of first four issues. It is through this unparalleled support and collaborative effort that the launch of this important publication became a reality to fulfill one of major goals of KAERA, "creating opportunities for and nurturing the environment of scholarly discourse, production, and collaboration among Korean-American and Korean researchers." I look forward to many years of enthusiastic scholarly exchange and cutting-edge discourses through this outlet, which will contribute to the advancement of educational conditions of Korean-American communities and beyond.

Sincerely,

Jae Hoon Lim

Jae Hoon Lim, Ph. D

2013-2014 KAERA President

KAERA Research Forum, *1(1),* 2-3

EDITORIAL

Dear Educational Researchers,

It is our pleasure to present to the educational community the compilation of six research papers in educational measurement to release the first *KAERA Research Forum*, an open access online scholarly forum, published by the Korean-American Educational Researchers Association. The *KAERA Research Forum* is a research report series that discusses a variety of topics in educational research and disseminates high-quality examples of theoretical and empirical research studies to inform the larger community of educational researchers and practitioners.

The purpose of the *KAERA Research Forum* is to create a venue of ideas by publishing and disseminating information about the most up-to-date scholarly endeavors and experiments pursued by members of the KAERA community and beyond. Taking the form of research briefs online, it aims to facilitate speedy and efficient sharing of new ideas among KAERA members and educational researchers at large. As such, the *KAERA Research Forum* provides an opportunity to both established scholars and emerging researchers. This inaugural issue of the *KAERA Research Forum* focuses on "Topics in Educational Measurement."

In this research forum, papers from the United States and South Korea that cover a wide array of research topics on the general theme of refining, understanding, and interpreting large-scale assessment data are presented. The collection of papers presents new theoretical methods and applications in improving how we analyze educational data, including discussions of educational significance for the measurement community and relevant policy implications. It is especially interesting to note that these six papers use various data from international, national, and state assessments, including:

- International assessment: the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA)
- National assessment in Korea: the College Scholastic Ability Test (CSAT) and the National Assessment of Educational Achievement (NAEA)
- Large-scale assessment in U.S.: Advanced Placement® (AP®)
- State assessment in U.S.: the Massachusetts Adult Proficiency Tests (MAPT)

The papers are presented in alphabetical order. In the first chapter, Youn-Jeng Choi, Natalia Alexeev, and Allan Cohen from the University of Georgia present an application of a covariate-based mixture item response theory model for explaining possible sources of differential item functioning using the Trends in International Mathematics and Science Study (TIMSS) data.

In the second chapter, HeeKyoung Kim at the Korea Institute for Curriculum and Evaluation (KICE) in South Korea demonstrates the use of a cognitive diagnostic model for developing achievement profiles of 6^{th}, 9^{th}, and 11^{th} grade students using the National Assessment of Educational Achievement (NAEA) data.

In the third chapter, YoungKoung Kim at The College Board, Lawrence DeCarlo at Teachers College, Columbia University, and Rosemary Reshetar at The College Board use the

Advanced Placement® (AP®) data to show the efficiency of a hierarchical rater model based on signal detection theory to adjust for rater differences over time when scoring constructed response items.

In the fourth chapter, Chanho Park at Keimyung University in South Korea, discusses the history, format, test development procedures, psychometric properties, and future trends of the College Scholastic Ability Test (CSAT), the only national college entrance examination used in South Korea.

In the fifth chapter, MinJeong Shin at the University of Massachusetts Amherst uses generalizability theory to investigate sources of error and to determine cost- and time-effective designs for optimizing standard setting studies using the Massachusetts Adult Proficiency Tests (MAPT) data.

Finally, in the sixth chapter, Ji Seung Yang at the University of Maryland and Li Cai at the University of California, Los Angeles present efficient estimation techniques for cross-level interactions in nonlinear multilevel latent variable models based on the Metropolis-Hastings Robbins-Monro algorithm through simulations and real-world data application using the Programme for International Student Assessment (PISA) data.

We hope that the first issue of the *KAERA Research Forum* presents opportunities for rich scholarly discussions and interactions that can further contribute to our community.

Won-Chan Lee, Editor
Yoon Soo Park, Associate Editor

KAERA Research Forum, *1(1)*, 4-14

RESEARCH ARTICLE

DIF Analysis using a Mixture 3PL Model with a Covariate on the TIMSS 2007 Mathematics Test

Youn-Jeng Choi, Natalia Alexeev, and Allan S. Cohen

University of Georgia

The purpose of this study was to explore what may be contributing to differences in performance in mathematics on the TIMSS 2007. This was done by using a mixture IRT modeling approach to first detect latent classes in the data and then to examine differences in performance on items taken by examinees in the different latent classes. An exploratory mixture 3pl model analysis detected two latent groups in the data. The model considered in this study used internet access as a covariate to illustrate the effect of the covariate on latent class membership.

Keywords: Differential item functioning, mixture IRT model, TIMSS 2007

The Trends in International Mathematics and Science Study (TIMSS) testing program is designed to provide information for countries to help them improve student learning in Mathematics and Science (Mullis, Martin, Ruddock, O'Sullivan, Arora, & Erberber, 2005). TIMSS reports average mathematics scale scores by country, thereby allowing comparisons among participating countries. In addition to scores, also of interest is how students in each country perform on each of the test items. One issue of concern, in this regard, is the extent to which the items on the tests perform the same in each country. One possible way to explain these differences would be to use differential item functioning (DIF).

There are a number of methods for detecting DIF based on comparisons among manifest groups (e.g., gender, ethnicity, or country) that are useful (Holland & Wainer, 1993; Thissen, Steinberg, & Wainer, 1988, 1993). Unfortunately, these methods may not easily explain what may be causing DIF. This is because the manifest group characteristics are typically only modestly associated with the cause of DIF. The purpose of this study is to explore what may be contributing to DIF among mathematics items taken in a subset of countries participating in the 4th grade TIMSS administered in 2007. We propose to use a method that leads to more information about the possible cause(s) of DIF.

THEORETICAL FRAMEWORK

Differential Item Functioning (DIF)

DIF arises when the item displays different statistical properties in different groups (Angoff, 1993). DIF is conditional, that is, it is defined as a differential propensity for a particular

response for examinees of the same ability but from different groups. It is typically observed when nuisance dimensions are in some way associated with manifest examinee characteristics (Ackerman, 1992; Roussos & Stout, 1996).

With standard DIF models, once a DIF item has been identified, unfortunately, little else is known about the students for whom the item functions differentially. This is because DIF is typically defined based on manifest group characteristics (e.g., gender, ethnicity) that are associated with, but do not explain, why examinees respond differentially to items (Cohen & Bolt, 2005).

Mixture 3-parameter Logistic Model

The Mixture 3-parameter Logistic Model (M3plM) is an extension of the mixture Rasch model (Rost, 1990), in which it is assumed that a population of examinees can be classified into a number of discrete latent classes. In the M3plM, the 3-parameter logistic model (3PL) is assumed to hold for each class but the item difficulty, discrimination, and guessing parameters may differ for the different classes. Each examinee is parameterized by an ability parameter (θ_j). The M3plM in Equation (1) associates a class membership parameter, *g*, with each item, *i*. The probability of a correct response in the M3plM is written as

$$P(y_{ij}=1|\theta_j)=\Sigma_{g=1}^{G}\pi_g\left[c_{ig}+(1-c_{ig})\frac{\exp(a_{ig}(\theta_j-b_{ig}))}{1+\exp(a_{ig}(\theta_j-b_{ig}))}\right], \qquad (1)$$

where g is an index for latent class, g =1,...,G, j =1,...,N examinees, θ_j is the latent ability of examinee j, π_g is the proportion of examinees for each class, a_{ig} is the discrimination parameter for item i in class g, b_{ig} is the difficulty parameter for item i in class g, and c_{ig} is the guessing parameter for item i in class g (Cohen & Bolt, 2005).

Mixture 3pl Model with Covariates

The M3plM can be extended to include covariates (M3plM-cov) that are useful in helping to explain why individuals can be classified in one of the latent classes (Smit, Kelderman, & van der Flier, 1999). There are at least two ways that covariates can be included into a model. If we consider the M3plM to be a multilevel model (e.g., Patz & Junker, 1999a, 1999b), then the covariates can be included as predictors at the latent class level. Another possibility is to incorporate the covariates so that they help to influence determining latent class membership. The latter approach was used in this study.

This particular M3plM-cov is a finite mixture regression model that incorporates covariates which can be applied to estimating latent class membership for an examinee or to explain the relationship between the means for an examinee on each of the G latent class means (Cho, Cohen, & Kim, 2008). Including the covariates in the same model to estimate the IRT model parameters as well as the mixing proportions makes it possible to diminish potential attenuation that can occur when covariates are not included in the model. In this study, the M3plM which included covariates was employed to estimate latent class memberships for each examinee. In

this formulation, examinees are classified as belonging to the latent class for which they have the highest probability of membership. The probability of a correct response in a M3plM-cov can be written as

$$P(y_{ij}=1|\theta_j)=\Sigma_{g=1}^{G}\pi_{jg|W_j}\left[c_{ig}+(1-c_{ig})\frac{\exp[a_{ig}(\theta_j-b_{ig})]}{1+\exp[a_{ig}(\theta_j-b_{ig})]}\right] \quad (2)$$

with

$$\pi_{jg|W_j}=\frac{\exp(\beta_{0g}+\Sigma_{p=1}^{P}\beta_{pg}W_{jp})}{\Sigma_{g=1}^{G}\exp(\beta_{0g}+\Sigma_{p=1}^{P}\beta_{pg}W_{jp})}, \quad (3)$$

where θ_j is the latent ability of examinee j, a_{ig} is the discrimination parameter for item i in class g, b_{ig} is the difficulty parameter for item i in class g, c_{ig} is the guessing parameter for item i in class g, and $P(G = g)= \pi_{jg}$ is the probability of examinee j belonging into class g. Group membership, g, has a multinomial distribution (Congdon, 2003) and the G latent groups are modeled as functions of the covariates W_{jp}, such that the π_{jg} is a multinomial logit regression. β_{pg} is a class-specific effect of covariate p on group membership. For identifiability, $\beta_{01} = 0$ and $\beta_{p1} = 0$ (Cho, Cohen, & Kim, 2008).

METHOD

Data

TIMSS 2007 consists of five sets of questions: an achievement test in mathematics, an achievement test in science, a student background questionnaire, a teacher background questionnaire (focusing on mathematics and science teaching), a school background questionnaire, and a curriculum questionnaire. The 11 multiple choice items and 15 short constructed response items (scored dichotomously) from the 2007 TIMSS 4th Grade math test were analyzed in this study (Foy & Olson, 2009). These items labeled as Unique ID measure three content domains: Data Display (4 items), Geometric Shapes and Measures (11 items), and Number (11 items).

A sample of seven countries was selected from the total sample participating in the 2007 TIMSS mathematics test for Grade 4. There were 1,845 4th Grade students in the following seven countries included in the sample: 356 students from Singapore, 273 students from Hong Kong, 264 students from Australia, 271 students from Austria, 302 students from Slovak Republic, 151 students from El Salvador, and 228 students from Qatar. These seven nations were chosen to have average scale scores that ranged from low to high on the test. Singapore and Hong Kong had the highest average mathematics scale scores (607 and 599, respectively) among the 36 countries which took the TIMSS 2007. Australia, Austria, and Slovak Republic had average mathematics scores (516, 505, and 496, respectively), and the lowest scores were for El Salvador and Qatar (330 and 296, respectively).

Estimation of Model Parameters

Estimation of model parameters was done using the Markov chain Monte Carlo (MCMC) estimation algorithm as implemented in the computer software WinBUGS (Spiegelhalter, Thomas, & Best, 2003). Three types of analyses are illustrated. First, an exploratory M3plM analysis was done using code written using the computer program WinBUGS to find how many latent groups exist in the data. Solutions for one to four latent classes were fit. The number of latent classes was determined using the Akaike's information criterion (AIC) and Bayesian information criterion (BIC). In cases where results from AIC and BIC differed, the BIC results were used to determine the number of latent classes as suggested by Li, Cohen, Kim, & Cho (2009). BIC values in Table 1 indicate that a model with two latent classes was the best fit to the data.

TABLE 1
Model Comparison Information Criteria for Mixture 3PL Solutions

Number of Classes	AIC	BIC
1	43370	43800
2	41740	42600
3	41460	42760
4	41250	42980

A DIF analysis was then conducted using the two-group solution suggested above. Examinees were classified using the modes of posterior densities for group membership into one of the two latent groups detected in the exploratory analysis. Then the data were analyzed using the computer program MULTILOG (Thissen, 2003). A two-group likelihood ratio test for DIF (Thissen, Steinberg, & Wainer, 1988, 1993) indicated that Items 2, 6, 7, 11, and 17 functioned the same in both latent classes (see Table 2). These items were used as anchor items, that is, they were constrained to be equal in both groups to anchor the metrics of the two classes. A M3plM-cov was next used to analyze the data. Item parameters of the unconstrained items were next compared.

Mixture 3-parameter IRT Model with a Covariate

A covariate was added to the M3plM for the purposes of helping determine group membership. The covariate used for an illustrative purpose in this study consisted of responses to a question about internet access available to examinees in the sample. This information was contained in the following question: “How many of these computers have access to the internet (email or World Wide Web) for educational purposes?” Answers to this question were coded as 1 = All, 2 = Most, 3 = Some, or 4 = None.

TABLE 2
DIF Tests Results for Class 1 and Class 2

			Class 1			Class 2		
Item No.	Unique ID	LR	a	b	c	a	b	c
1	M031235	48.6	0.67	0.23	0.00	1.03	-0.62	0.04
2	M031285	1.1*	0.92	0.24	0.14	0.72	0.15	0.00
3	M031050	13.4	1.27	0.01	0.36	2.10	-0.46	0.37
4	M031258	24.6	1.49	0.72	0.37	1.18	-0.46	0.00
5	M031334	29.5	2.53	0.59	0.44	1.06	-0.37	0.27
6	M031255	6.0*	2.44	0.08	0.65	0.89	-0.85	0.30
7	M031041	6.4*	1.13	0.21	0.54	0.89	-0.72	0.00
8	M031350A	294.5	0.98	0.16	0.00	2.49	-1.12	0.00
9	M031350B	242.5	1.12	-0.27	0.00	3.34	-1.32	0.06
10	M031350C	131.2	0.58	0.41	0.00	1.19	-0.89	0.00
11	M031274	6.9*	1.68	0.02	0.88	0.95	-1.60	0.00
12	M031240	10.0	1.67	0.72	0.78	0.69	-1.39	0.00
13	M041052	39.0	2.18	0.26	0.73	1.53	-1.32	0.45
14	M041056	62.8	0.98	-0.60	0.31	1.83	0.08	0.15
15	M041069	207.7	1.18	0.03	0.48	0.47	4.11	0.07
16	M041076	119.0	0.38	-1.42	0.60	0.47	0.05	0.00
17	M041281	7.3*	0.76	-0.81	0.60	1.25	-1.41	0.10
18	M041164	83.0	0.21	-12.88	0.31	0.47	-1.51	0.19
19	M041146	43.3	1.17	-0.01	0.56	1.37	-1.35	0.04
20	M041152	139.2	0.67	-0.02	0.75	0.67	0.52	0.14
21	M041258A	15.3	0.83	-1.09	0.01	1.04	-1.39	0.00
22	M041258B	82.1	0.55	0.20	0.00	0.54	2.71	0.00
23	M041131	13.3	0.49	0.90	0.00	0.53	0.45	0.10
24	M041275	12.1	0.57	-2.43	0.00	0.70	-1.39	0.00
25	M041186	47.8	0.41	-1.37	0.00	1.22	-1.13	0.04
26	M041336	18.3	1.13	-0.08	0.05	0.37	-0.23	0.00

Note. * $p(\chi^2(df = 3) < 7.81) < .05$

The probability of a correct response in a M3plM-cov is given in Equation (2). Following was used for each level of the internet access information: β_{0g} indicates "none," β_{1g} indicates "all," β_{2g} indicates "most," and β_{3g} indicates "some." The coefficients for each level of the covariate were set at 0 in Class 1 for identification. That is, $\beta_{11} = \beta_{21} = \beta_{31} = 0$. For identifiability, β_{01} was also set to 0.

RESULTS

Estimation of M3plM with a Covariate

Table 3 shows the item parameter estimates for discrimination, difficulty, and guessing. The values of Items 2, 6, 7, 11, and 17 were fixed to be the same, respectively, to anchor the metrics of the two latent classes so each estimated item parameter was on the same metric in the two latent classes.

TABLE 3

Item Parameters for Two-Class Solution for Model with Internet-access Covariate

		Class 1			Class 2		
Item No.	Unique ID	a	b	c	a	b	c
1	M031235	1.11	0.78	0.17	1.91	-0.43	0.07
2	M031285	0.85	0.15	0.00	0.85	0.15	0.00
3	M031050	1.62	-0.16	0.21	1.50	-0.42	0.30
4	M031258	1.36	0.48	0.15	1.93	-0.22	0.03
5	M031334	2.15	0.63	0.36	1.24	-0.10	0.22
6	M031255	1.32	-0.48	0.35	1.32	-0.48	0.35
7	M031041	0.94	-0.54	0.00	0.94	-0.54	0.00
8	M031350A	1.19	0.31	0.15	4.72	-1.07	0.02
9	M031350B	1.21	-0.28	0.25	6.39	-1.30	0.06
10	M031350C	0.94	0.97	0.17	2.18	-0.76	0.03
11	M031274	1.13	-1.34	0.00	1.13	-1.34	0.00
12	M031240	0.65	-1.85	0.30	1.19	-0.72	0.12
13	M041052	1.28	-1.11	0.23	1.61	-1.89	0.27
14	M041056	1.98	-0.36	0.15	2.07	0.60	0.14
15	M041069	1.80	-0.10	0.14	1.67	2.51	0.07
16	M041076	1.00	-1.26	0.25	1.34	0.74	0.09
17	M041281	1.04	-1.24	0.03	1.04	-1.24	0.03
18	M041164	1.25	-2.99	0.26	0.86	-0.55	0.29
19	M041146	1.19	-0.86	0.24	2.16	-1.19	0.06
20	M041152	1.43	-0.69	0.19	1.24	0.79	0.14
21	M041258A	1.01	-1.10	0.28	1.86	-1.07	0.07
22	M041258B	1.27	0.72	0.08	1.55	2.55	0.02
23	M041131	0.91	1.67	0.16	0.97	0.78	0.13
24	M041275	0.85	-2.19	0.27	1.36	-0.77	0.12
25	M041186	0.76	-0.66	0.23	1.95	-0.90	0.07
26	M041336	1.86	0.28	0.20	0.88	0.11	0.11

Note. a = discrimination parameter, b = difficulty parameter, c = guessing parameter

The sample consisted of two high-performing countries (Hong Kong & Singapore), two low-performing countries (El Salvador & Qatar), and three average performing countries (Australia, Austria, & Slovak Republic). The two latent classes appeared to be formed mainly along a dimension reflecting differences in ability. Membership in Class 1, in other words, included more examinees from high-performing countries and fewer from low-performing countries. Members of Class 2 tended to have more examinees from low-performing countries. Almost all students from Hong Kong (n = 269) and Singapore (n = 319), for example, belonged to Class 1; 224 examinees from Qatar and 141 examinees from El Salvador belonged Class 2 (see Table 4).

TABLE 4
Latent Classes Make-up by Country for Model with Internet-access Covariate

Country	Class 1	Class 2	Total
Australia	159 (8.6%)	105 (5.7%)	264 (14.3%)
Austria	100 (5.4%)	171 (9.3%)	271 (14.7%)
El Salvador	10 (0.5%)	141 (7.6%)	151 (8.2%)
Hong Kong	269 (14.6%)	4 (0.2%)	273 (14.8%)
Qatar	4 (0.2%)	224 (12.1%)	228 (12.4%)
Singapore	319 (17.3%)	37 (2.0%)	356 (19.3%)
Slovak Republic	55 (3.0%)	247 (13.4%)	302 (16.4%)
Total	916 (49.6%)	929 (50.4%)	1845 (100.0%)

Further examination of results, however, suggested that there was more than just performance defining latent class membership. As is noted in the sequel, items that appeared to function differently between latent classes were ones that differed in either content topic or problem type or both.

TABLE 5
Coefficients for Model with Internet-access Covariate

Class	All	Most	Some	None
1	0	0	0	0
2	-2.434	-1.632	-1.017	1.829

The use of a covariate was illustrated in this study with information that described internet access available to examinees in the sample. Coefficients for the covariate were compared between the two classes to examine the effect of the covariate on latent class membership.

Results suggest that as the coefficient for the covariate in Class 2 increased, the

frequency of internet access decreased (see Table 5). This means that the members of Class 2 had less internet access than members of Class 1. Table 6 shows the frequencies and proportions for each level of the covariate for each latent class. It can be seen clearly that more members of Class 1 (N = 701) had "All" internet access than did members of Class 2 (N = 476). More members of Class 2 had "None" (N = 170) than members of Class 1. A chi-square analysis indicated that internet access was associated with class membership ($p < .001$)

TABLE 6
Internet Access Composition in Latent Classes

Class	All	Most	Some	None	Total
1	701 (41.8%)	94 (5.6%)	30 (1.8%)	27 (1.6%)	858 (50.8%)
2	476 (28.4%)	122 (7.3%)	56 (3.3%)	170 (10.1%)	824 (49.2%)
Total	1177 (70.2%)	216 (12.9%)	86 (5.1%)	197 (11.8%)	1676 (100.0%)

Comparison of Latent Groups and Item Performance

The group mean theta values of Class 1 and Class 2 were 0 and −1.58, respectively. This indicates that Class 2 was lower in ability as measured by the TIMSS 2007. The proportions of examinees classified into Classes 1 and 2 were 49.6 (n = 916) and 50.4 (n = 929), respectively. Differences in item performance suggest that group membership was indicative of more than just differences in ability. Members of Class 1 performed better on Items 14, 15, and 16. These were the only items on TIMSS 2007 that involved fractions (see Table 7). One possible explanation for this pattern between latent classes is that the differences in performance on fractions items may reflect instructional sequencing. If fractions are just being introduced, for example, one may see a pattern of performance such as that seen here between low-performing and high-performing countries.

Two items (Items 12 and 24) measured reading and understanding of data also appeared to operate differently in both groups. These items (Items 12 and 24) focused on reading the data and compiling a new table or a graph. More advanced items (Items 18, 20, and 22) measured concepts dealing with geometric shapes. Members of Class 1 did better on these. This could be another example of instructional sequencing differences that might be reflected in differences between the two latent classes.

There were a total of eight items (Items 1, 4, 5, 8, 9, 10, 13, and 23) on which members of Class 2 did better than expected, given that this class was of lower ability. Although members of this class did not perform better than examinees in Class 1, their performances were more similar to that of members of Class 1. Two of these items (Items 4 and 5) measured concepts related to patterns, and could be considered as pre-algebra items. Both classes mastered this concept almost at the same level. With the exception of Item 13, which describes place value, the remainder of these eight items all items on which members of Class 2 did better were actually more difficult for examinees in both classes. This may be an explanation of better than expected performance on these items for members of Class 2.

TABLE 7
Items that Appeared to Function Differently in the Two Classes

Item	Unique ID	Content Domain	Cognitive Doman	Description	Discrimination		Difficulty	
					Class 1	Class 2	Class 1	Class 2
Benefiting Class 1								
12	M031240	Data display	Applying	Recompiling data	0.65	1.19	-1.85	-0.72
14	M041056	Number	Knowing	Concept of a fraction	1.98	2.07	-0.36	0.60
15	M041069	Number	Knowing	Equal fractions	1.80	1.67	-0.10	2.51
16	M041076	Number	Knowing	Adding fractions	1.00	1.34	-1.26	0.74
18	M041164	Geom. shapes & measures	Knowing	Understanding Symmetry	1.25	0.86	-2.99	-0.55
20	M041152	Geom. shapes & measures	Applying	Area of a rectangle	1.43	1.24	-0.69	0.79
22	M041258B	Geom. shapes & measures	Reasoning	Describing triangles	1.27	1.55	0.72	2.55
24	M041275	Data display	Applying	Recompiling data	0.85	1.36	-2.19	-0.77
Benefiting Class 2								
1	M031235	Number	Reasoning	Understanding Multiples	1.11	1.91	0.78	-0.43
4	M031258	Number	Reasoning	Finding pattern rule	1.36	1.93	0.48	-0.22
5	M031334	Number	Applying	Recognizing pattern rule	2.15	1.24	0.63	-0.10
8	M031350A	Geom. shapes & measures	Applying	Measure distances	1.19	4.72	0.31	-1.07
9	M031350B	Geom. shapes & measures	Reasoning	Measure distances	1.21	6.39	-0.28	-1.30
10	M031350C	Geom. shapes & measures	Applying	Measure distances	0.94	2.18	0.97	-0.76
13	M041052	Number	Knowing	Place value	1.28	1.61	-1.11	-1.89
23	M041131	Geom. shapes & measures	Knowing	Scaling	0.91	0.97	1.67	0.78

DISCUSSION

The two latent groups reflect secondary nuisance dimensions in the data that were not accounted for in the 3PL model and that are potential causes of DIF in test items. The two latent classes appeared to be formed mainly along a performance dimension. One problem with the usual analysis of DIF among manifest groups is that differences between such groups are usually only modestly associated with the causes of DIF. The advantage of the mixture IRT modeling approach is that the characteristics of members of latent classes detected in the sample and of the item characteristics that are differentially harder or easier for particular latent classes may help lead to a more direct interpretation of the possible causes of DIF.

Including covariates in the model has been shown to have the potential to improve the detection of latent classes as well as the interpretation of differences among these groups. The availability of internet access in schools appeared to be associated with differences in performance between the latent classes.

Results indicated that differential performances between the latent classes were reflected largely in differences in specific items. Members of Class 1, for example, performed better on items in TIMSS 2007 that involved fractions, as well as on more advanced items measuring geometric shapes. One conjecture is that these differences may reflect differences in curricula between the countries.

Members of Class 2, the lower ability class, had less internet access. There were total eight items (Item 1, 4, 5, 8, 9, 10, 13, and 23) that members of Class 2 did better than expected given that this class was of lower ability. Examinees in this class did not perform better than examinees in Class 1, but their performances were more similar to members of Class 1.

The results from this study are helpful for understanding how differences may arise between countries although only seven of the 36 countries that took the TIMSS 2007 were examined. A mixture modeling approach was not done on all the countries in the TIMSS 2007 dataset. Inclusion of the full sample may point to other characteristics of the latent classes or even other latent group classifications that would be helpful in understanding differences in performance among countries. It would have been preferable to use a single algorithm for all the analyses. In this study, MULTILOG does not handle detection of latent classes, but is generally much faster than doing an MCMC analysis. For purposes of expediency, therefore, two estimation algorithms were used. This will be corrected in future work on this problem.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Angoff, W. H. (1993). Perspectives in differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cho, S.-J., Cohen, A. S., & Kim, S-H. (2008). *Markov chain Monte Carlo estimation of a mixture Rasch model.* Paper presented at the International Meeting of the Psychometric Society, Montreal, CA.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133-148.

Congdon, P. (2003). *Applied Bayesian modeling*. New York: John Wiley.

Foy, P., & Olson, J. F. (2009). *TIMSS 2007 user guide for the international database.* Chestnut Hill, MA: Boston College.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for dichotomous mixture IRT models. *Applied Psychological Measurement, 33*, 353-373.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: Boston College.

Patz, R. J., & Junker, B.W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

Patz, R. J., & Junker, B.W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

Smit, A., Kelderman, H., & van der Flier, H. (1999). Collateral information and mixed Rasch models. *Methods of Psychological Research Online, 4*.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS (version 1.4) [computer program]. Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health.

Thissen, D. (2003). *MULTILOG for Windows* [computer program]. Lincolnwood, IL: Scientific Software International, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates *Behavioral Statistics, 24*, 342-366.

KAERA Research Forum, *1(1)*, 15-25

RESEARCH ARTICLE

Application of Cognitive Diagnostic Model for Achievement Profile Analysis

HeeKyoung Kim

Korea Institute for Curriculum and Evaluation

This study empirically verified the possibility of analyzing students' achievement profile information by applying a CDM to a large-scale assessment dataset with more than 600,000 examinees. Also, characteristics of achievement profiles of Korean students' Korean Language Arts and Mathematics were analyzed. Characteristics of achievement profiles were compared across grade levels, gender, and regional characteristics of schools. Comparison was also made for the multi-cultural background of families, reflecting the increased social interests and needs for educational support. The results of this study show that a CDM could serve as an appropriate technical solution to the needs that assessment results should provide information that enhances student learning and facilitate communications with teachers and parents to support educational improvement.

Keywords: Cognitive Diagnostic Model (CDM), achievement profile, large-scale assessment

The National Assessment of Educational Achievement (NAEA) in Korea is a criterion-referenced assessment built around a framework based on the national curriculum in subject areas such as Korean Language Arts and Mathematics. The NAEA was developed to measure student achievement, to diagnose proficiency of each student, to provide information for improving teaching and learning practices, and to give evidence that could support educational policies for improving effective education. However, under the current assessment design of the NAEA, test results are provided to students as a form of a single overall score or a performance descriptor, which often fails to provide fine-grained information about strengths and weaknesses of individual students. For this reason, we seek for innovative ways of analyzing test results using a Cognitive Diagnostic Model (CDM) in such a way that individual students can fully benefit from the assessment by getting detailed feedback.

This study has two main objectives. First, previous studies on CDMs were performed based on sample data. Little research has been carried out using data from an entire population. Because the NAEA is a census test, the total number of examinees of the NAEA is approximately 600,000 per each participating grade level. Therefore, the feasibility of CDM implementation to accommodate extremely large number of examinees will be explored using NAEA data. Second, a CDM will be applied to the assessment results of the NAEA in order to

provide academic characteristics of Korean students. CDMs could serve as an appropriate technical solution to the needs that assessment results should provide information that enhances student learning and facilitates communications with teachers and parents to support educational improvement.

METHOD

This study used data from the NAEA administration to 6th, 9th, 11th graders in 2011. Sample sizes are presented in Table 1. At the high school level (11th graders), only students attending college-bound high schools are included in the analyses.

TABLE 1
Sample Size for Each Grade

Subject	Sample Size		
	6th Grade	9th Grade	11th Grade
Korean Language Arts	578,020	628,200	486,675
Mathematics	576,166	628,369	469,797

Cognitive attributes measured by Korean Language Arts and Mathematics tests of the NAEA 2011 were identified, and Q-matrices were developed to connect items to relevant attributes[1]. Also, validation procedures based on statistical analyses and CDM-based item parameters were used for detecting misspecification in Q-matrices. Multiple regression analyses were performed and Jaccard Index[2] was calculated as a statistical approach and item parameter estimates from the Fusion model were also reviewed to validate the Q-matrices. Tables 2 and 3 summarize the item compositions for each cognitive attribute, which shows the final version of Q-matrices.

[1] A total of 8 and 5 content experts in Korean Language Arts and Mathematics, respectively, participated in the process of Q-matrix construction.

[2] The index, also known as the Jaccard similarity coefficient, was first developed by Jaccard(1901). The Jaccard index is defined as $J(A,B) = |A \cap B| / |A \cup B|$ (as cited in Hannig, 2004).

TABLE 2

Item Compositions for Each Cognitive Attributes (Korean Language Arts)

Cognitive Attributes	6th	9th	11th
1. Knowledge of Korean Language Arts	9, 20, CR3	7, 10, 21, 25, CR3	8, 9, 11, CR1, CR2, CR3
2.Understanding the meaning of a word	15, CR4, CR5	5, 21, CR3	16, CR3, CR4, CR6
3.Analyze the connection of sentences	7, 10	6, 14, 21, 28, CR2	10, 24, CR3, CR6
4.Understand specific information	3, 4, 14, 25, 26	2, 9, 16, 18, CR1	1, 2, 3, 15, 22
5.Understand the main idea	CR1	3, CR5	1, 4, 12, 18, CR3, CR4
6.Infer omitted information	1, 6, 13, 22, CR2	15, CR3	13, CR5
7.Infer implicit information	11, 12, 19, 24	10, 23, 24, 26, 27, CR6	2, 19, 23, 26, 27, 28, 29, 30, CR5
8.Understand structure and development of the text	17, 27	4, 13, 17, CR4, CR6	7, 25, CR1, CR4
9.Analyze intention and view of the author or the speaker	-	1, 11, 19, 22	17
10.Explore and apply	-	-	11
11.Evaluate appropriateness and credibility of the content	2, 8, 21	8, 12, 20	5, 6, 15, 20
12. Evaluate appropriateness of the structure or expressions	5, 18	-	14, 21, 25
13. Rhetorical knowledge	16, 23, CR5	CR4	5
14. Express and revise the draft	CR1, CR2, CR5	28, CR3, CR5	15, CR2, CR3, CR6

TABLE3

Item Compositions for Each Cognitive Attributes (Mathematics)

Cognitive Attributes	6th	9th	11th
1. Simple arithmetic operations	1, 5, 6, 10, 13, 14, 15, 17, 19, 21, 24, 25, CR1, CR2	2, 6, 18, CR1	1, CR1, CR2
2. Routine algebraic procedures	7	3, 4, 9, 10, 14, 15, 16, 17, 19, 20, 23, 26, 27, CR2, CR4	2, 4, 6, 11, 13, 14, 16, 17, 18, 19, 20, 21, 22, 24, 25, CR3
3. Understanding principles and rules	3, 12, 13, 21	18, 21, 24, 25, 26, 28	2, 3, 4, 9, 11, 12, 14, 16, 19, 22, 25, 26, 27, 28, CR2, CR4
4. Understanding concepts and properties	1, 2, 4, 5, 6, 9, 10, 11, 14, 15, 17, 22, 24, 25, CR1, CR2, CR3	1, 3, 5, 6, 7, 8, 11, 13, 14, 19, 22, CR1, CR2, CR3	1, 5, 7, 8, 10, 13, 18, CR1
5. Analyzing	16, CR4	-	15, 20, 23, 24, 26, 29, CR3
6. Inductive reasoning	19, 23, CR4	-	-
7. Inductive deducing	9, 15, 16	-	-
8. Deductive justifying	-	22, CR3	CR3
9. Representing using picture, table, graph, formula, symbol, writing, etc.	8, 16, 18, 19, CR2, CR3, CR4	10, 11, 20, 23, 27, CR4	10, 12, 14, 17, 20, 21, 23, 24, 25, 26, 29, CR4
10. Analyzing context in problem solving	13, 14, 15, 21, 23, 25, CR4	23, 27	10, 21, 27
11. Analyzing data on functions and equations of figures	18	12, 15, CR2	8, 12, CR4
12. Analyzing statistical data	6, 7, 8, 10, 20, 23	18, 29	-
13. Analyzing information from figures and diagrams	3, 4, 17, 22	5, 17, 19, 21, 22, 24, 25, 26, 28, CR1, CR3, CR4	21, 28, 29

There is not much CDM software currently available to accommodate polytomously-scored items. The Fusion model has been recently extended to accommodate polytomously-scored items (Fu, 2005). Since the NAEA is a mixed format test consisting of both multiple-choice and constructed-response items, this study used currently available software for the extended Fusion model, using Arpeggio version 3.1 (Dibello & Stout, 2010).

RESULTS

This study empirically verified the possibility of producing students' attribute mastery profile information by applying a CDM to large-scale data of more than 600,000 examinees. The CDM was also applied to analyze characteristics of achievement profiles that Korean students show on Korean Language Arts and Mathematics. Characteristics of achievement profiles were compared across grade levels, gender, and regional characteristics of schools. Comparison was also made for the multi-cultural background of families, reflecting the increased social interests and needs for educational support.

TABLE4
Classification of Consistency of Mastery/Non-mastery in Fusion Model
(Korean Language Arts)

Cognitive Attributes	6th		9th		11th	
	CCR	TRC	CCR	TRC	CCR	TRC
1. Knowledge of Korean Language Arts	0.91	0.85	0.91	0.84	0.95	0.91
2. Understanding the meaning of a word	0.89	0.81	0.90	0.82	0.94	0.90
3.Analyze the connection of sentences	0.90	0.84	0.89	0.81	0.92	0.86
4. Understand specific information	0.90	0.83	0.93	0.88	0.94	0.89
5. Understand the main idea	0.81	0.73	0.87	0.79	0.95	0.90
6. Infer omitted information	0.89	0.81	0.84	0.74	0.93	0.87
7. Infer implicit information	0.91	0.85	0.95	0.90	0.94	0.89
8. Understand structure and development of the text	0.88	0.80	0.93	0.88	0.87	0.77
9. Analyze intention and view of the author or the speaker	1.00	-	0.99	0.97	0.94	0.89
10. Explore and apply	1.00	-	1.00	-	0.89	0.81
11. Evaluate appropriateness and credibility of the content	0.89	0.82	0.91	0.85	0.87	0.77
12. Evaluate appropriateness of the structure or expressions	0.88	0.79	1.00	-	0.88	0.80
13. Rhetorical knowledge	0.87	0.78	0.86	0.78	0.91	0.85
14. Express and revise the draft	0.92	0.86	0.92	0.86	0.91	0.86
Average	0.90	0.81	0.92	0.84	0.92	0.86

TABLE 5

Classification Consistency of Mastery/ Non-mastery in Fusion Model (Mathematics)

Cognitive Attributes	6th		9th		11th	
	CCR	TRC	CCR	TRC	CCR	TRC
1. Simple arithmetic operations	0.89	0.81	0.93	0.87	0.97	0.95
2. Routine algebraic procedures	0.90	0.83	0.98	0.96	0.96	0.93
3. Understanding principles and rules	0.82	0.74	0.90	0.82	0.97	0.95
4. Understanding concepts and properties	0.97	0.95	0.96	0.92	0.95	0.90
5. Analyzing	0.88	0.81	-	-	0.88	0.80
6. Inductive reasoning	0.88	0.82	-	-	-	-
7. Inductive deducing	0.86	0.76	-	-	-	-
8. Deductive justifying	-	-	0.91	0.84	0.89	0.80
9. Representing using picture, table, graph, formula, symbol, writing, etc.	0.91	0.85	0.94	0.90	0.94	0.89
10. Analyzing context in problem solving	0.89	0.82	0.88	0.79	0.85	0.75
11. Analyzing data on functions and equations of figures	0.95	0.92	0.90	0.82	0.89	0.80
12. Analyzing statistical data	0.89	0.82	0.88	0.78	-	-
13. Analyzing information from figures and diagrams	0.89	0.83	0.93	0.87	0.85	0.76
Average	0.89	0.83	0.92	0.86	0.92	0.85

An examinee parameter estimated from a CDM represents the posterior probability of mastery (PPM) meaning the probability that an examinee masters a particular cognitive attribute. An examinee was classified as mastery status if the PPM was above .5, otherwise as non-mastery in this study.

In order to evaluate the model-data fit of the Fusion model, a simulation study based on 100,000 simulees was conducted to obtain classification consistency indices between mastery/non-mastery decisions for each attribute. Tables 4 and 5 represent the proportion of accurately classified examinees. That is, the CCR (Correct Classification Rate) means the consistency between the true mastery status and the estimated mastery status among 100,000 simulees. The TCR (Test-Retest Consistency) represents probability of consistent mastery decisions when the test was assumed to be administered for the same examinees repeatedly.

Tables 6 and 7 present the attribute mastery proportion of the NAEA 2011 Korean Language Arts and Mathematics for each grade.

TABLE 6

Attribute Mastery/Non-mastery Proportion of the NAEA 2011 Korean Language Arts (%)

Cognitive Attributes	6th		9th		11th	
	Master	Non-master	Master	Non-master	Master	Non-master
1. Knowledge of Korean Language Arts	76.17	23.83	74.44	25.56	81.24	18.76
2.Understanding the meaning of a word	69.35	30.65	81.10	18.90	80.31	19.69
3.Analyze the connection of sentences	76.63	23.37	65.62	34.38	72.50	27.50
4.Understand specific information	68.37	31.63	69.88	30.12	71.02	28.98
5.Understand the main idea	72.84	27.16	71.41	28.59	75.31	24.69
6.Infer omitted information	68.00	32.00	65.25	34.75	79.64	20.36
7. Infer implicit information	73.01	26.99	66.55	33.45	56.01	43.99
8. Understand structure and development of the text	65.16	34.84	71.87	28.13	60.67	39.33
9. Analyze intention and view of the author or the speaker	-	-	95.43	4.57	74.83	25.17
10. Explore and apply	-	-	-	-	72.55	27.45
11. Evaluate appropriateness and credibility of the content	70.82	29.18	66.27	33.73	59.82	40.18
12. Evaluate appropriateness of the structure or expressions	64.74	35.26	-	-	62.74	37.26
13. Rhetorical knowledge	66.19	33.81	79.36	20.64	82.24	17.76
14. Express and revise the draft	80.15	19.85	82.20	17.80	83.49	16.51

TABLE 7

Attribute Mastery/ Non-mastery Proportion of the NAEA 2011 Mathematics (%)

Cognitive Attributes	6th		9th		11th	
	Master	Non-master	Master	Non-master	Master	Non-master
1. Simple arithmetic operations	47.78	52.22	64.24	35.76	66.61	33.39
2.Routine algebraic procedures	74.52	25.48	64.56	35.44	41.04	58.96
3.Understanding principles and rules	73.18	26.82	39.91	60.09	62.29	37.71
4.Understanding concepts and properties	76.70	23.30	52.85	47.15	46.18	53.82
5. Analyzing	81.53	18.47	-	-	59.59	40.41
6. Inductive reasoning	84.19	15.81	-	-	-	-
7. Inductive deducing	53.19	46.81	-	-	-	-
8. Deductive justifying	-	-	69.78	30.22	42.25	57.75
9. Representing using picture, table, graph, formula, symbol, writing, etc.	77.58	22.42	34.55	65.45	26.27	73.73
10. Analyzing context in problem solving	82.35	17.65	52.52	47.48	40.40	59.60
11. Analyzing data on functions and equations of figures	83.18	16.82	61.13	38.87	48.10	51.90
12. Analyzing statistical data	73.52	26.48	51.03	48.97	-	-
13. Analyzing information from figures and diagrams	83.45	16.55	56.73	43.27	34.89	65.11

As grade level increased from 6th, 9th, to 11th, the proportion of students who mastered Attribute 7 (Infer implicit information) decreased from 73%, 67%, to 56% in Korean Language Arts; while Attribute 9 (Representing using picture, table, graph, formula, symbol, writing, etc.) decreased with even faster rate from 78%, 35%, to 26% in Mathematics. As grade level increased, students appeared to feel more difficult in reading a passage and making inferences about what they read. In addition, the findings also imply that effective learning strategies should be developed to enhance learning mathematical representations such as tables and graphs that students feel more difficult as the grade level increases.

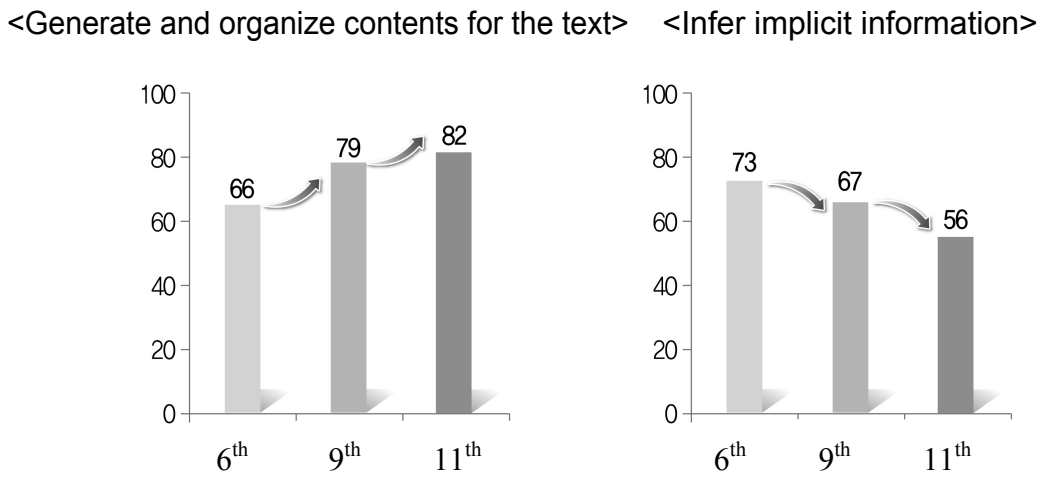

Figure 1. Achievement Profile Pattern by Grade Level

Gender difference in Korean Language Arts was found for the mastery of Attribute 7(Infer implicit information), with consistent out-performance of female students than males by approximately 14% difference across grades. In Mathematics, the mastery proportion of female students was generally higher than that of male students for the 6th and 9th grades, while opposite patterns were observed for the 11th grade. The greatest gender difference for the 11th grade was observed in Attribute 9 (Representing using picture, table, graph, formula, symbol, writing, etc.).

When the mastery proportion was compared among regional characteristics, students in the rural area showed the lowest proportion of mastery in Attribute 6 (Infer omitted information) of Korean Language Arts compared to other types of regions (large city and small/medium city) for all grades. In Mathematics, students in the rural area showed the lowest proportion of mastery in Attribute 3 (Understanding principles and rules) for the 9th and 11th grades.

Recently, the number of students from multi-cultural families has increased and is expected to continue to grow in South Korea due to rising cases of international marriage and influx of foreign workers. Identifying difficulties that they might have due to differences in language and culture is important in enhancing the overall academic achievement of our country. In Korean Language Arts, gaps in the mastery proportion between students from multi-cultural families and those from typical Korean families appeared to decrease (12%p → 9%p → 8%p) as grade level increased (6th → 9th → 11th). However, a closer look at the four types of multi-cultural families (students from foreign families, students from international marriage, immigrant adolescents, and North Korean refugees) revealed that the decreased gaps were mainly due to relatively better performance of students from foreign families and from international marriage and that learning deficiency of immigrant adolescents and North Korean adolescent refugees was getting worse as grade level increased. Therefore, educational support for these two types of multi-cultural families seems to be imperative.

DISCUSSION AND CONCLUSIONS

This study applied a CDM to analyze test results of Korean Language Arts and Mathematics of the NAEA 2011. Teachers should be able to gain information from the academic achievement profiles of each student in their classroom about whether students are learning in a balanced way. Information on achievement profiles would provide an effective solution to support students' academic growth rather than only focusing on the overall test score, because it gives a better understanding about various aspects of student achievement.

It should be noted from this study that, as grade level increased from 6th, 9th, to 11th, the proportion of students who mastered Attribute 13 (generate and organize contents for the text) increased from 66%, 79%, to 82% in Korean Language Arts; while that of Attribute 7 (infer implicit information) decreased from 73%, 67%, to 56% in Mathematics (see Figure 1). The increasing pattern of the mastery proportion of Attribute 13 (generates and organizes contents for the text) can be understood as a natural phenomenon of student development. Since students feel a high level of cognitive burden in the process of translating what they think into writing, the cognitive ability needed for expression seems to increase as grade level increases.

However, the decreasing pattern of Attribute 7 (Infer implicit information) implies that appropriate reading comprehension skills are not developed in a systematic way although contents of passages become more sophisticated as grade level increases. It seems to be a common practice in the Korean language arts education that more focus is given to understanding factual knowledge in a text, while less focus is given to making inference about implicit information such as values of the author or social backgrounds of the text. Therefore, in order to improve students' inferential skills, effective strategies supplementing the current practice of reading instruction are needed in such a way that students can explore diverse expressions that give insights about the author, social backgrounds, and contexts.

In Mathematics, the mastery proportion of Attribute 9 (Representing using picture, table, graph, formula, symbol, writing, etc.) decreased with an even faster rate (78% → 35% → 26%) as grade level increased. Attribute 9 refers to the "mathematical communication skills" which is closely tied to the problem solving ability that is given more attention recently. Problem solving ability gives more emphasis on the process to obtain a correct answer through discussion or communication with others, rather than on the problem solving itself. Therefore, instructional strategies need to be developed in such a way that they encompass various activities involving mathematical representations and give less time and effort in practice and drills for solving problems.

Analyzing large scale test data with more than 600,000 examinees based on a CDM approach is the first attempt that has been made in South Korea. It turned out that the time spent for dealing with more than 600,000 examinees using CDM software with a mathematically complex structure algorithm was approximately 2~3 days per subject in each grade. If we take into account the time needed to prepare for reporting the CDM-based achievement profile information to students, analysis time needs to be reduced if possible. Also, in school settings, teachers might want to evaluate individual students' mastery level of core cognitive attributes that are taught at the end of each unit or each month. In order to make a closer link between results from such tests to instruction, practical ways or software for applying CDMs in school settings need to be considered and studied

REFERENCES

Dibello, L., & Stout, W. (2010). *Arpeggio version 3.1* [Computer Program]. Chicago: Applied Informative Assessment Research Enterprises.

Fu, J. (2005). A *polytomous extension of the Fusion Model and its Bayesian parameter estimation*. Unpublished doctoral dissertation. University of Wisconsin-Madison.

Hannig, C. (2004). *Cluster-wise assessment of cluster stability.* Unpublished research report, Department of Statistical Science, University of College London.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des Dranses et dans quelques règions voisines. *Bulletin de la Sociėtė Vaudoise des Sciences Naturelles, 37,* 241-272

KAERA Research Forum, *1(1)*, 26-35

RESEARCH ARTICLE

Linking with Constructed Response Items: A Hierarchical Model Approach with AP Data

YoungKoung Kim
The College Board

Lawrence T. DeCarlo
Teachers College, Columbia University

Rosemary Reshetar
The College Board

Linking with constructed response (CR) items is more complex than linking with multiple choice items, because CR items are scored by raters, which introduces another source of variation into the scores. When an item response theory model is used as a rater model, one must include a condition to link raters, as has been previously discussed (e.g., Tate, 1999). Here it is shown that, when a hierarchical-rater signal-detection model is used, the CR item parameters are not distorted due to changes in the rater parameters (e.g., severity), and thus a rater-linking condition is not needed. The approach is illustrated with an application to data from the Advanced Placement® (AP®) Studio Art portfolios 3-D Design.

Keywords: Rater effect, Linking, Trend Scoring, Hierarchical-Rater Signal-Detection (HRM-SDT) and Item Response Theory (IRT)

When test forms consisting of constructed response (CR) items are linked, adjusting for changes in the raters' scoring over time needs to be considered, in addition to adjusting for changes in examinees and in test form difficulty. Studies have found that traditional linking methods for CR items can lead to inaccurate results because changes in rater scoring over time are confounded with changes in examinee ability (Tate, 1999, 2000, 2003; Kamata & Tate, 2005; Kim, Walker, & McHale, 2010a, 2010b). For this reason, methods for "rater linking" or "trend scoring," where some of the raters from the second scoring occasion also score CR items from the first scoring occasion (referred to from here on as Time 1 and Time 2), have been developed.

Tate (1999, 2000) suggested a procedure within an IRT framework to link rater scores across Time 1 and Time 2, which is referred to as the *IRT trend scoring approach* in the present study. The IRT trend scoring approach requires that a group of raters in Time 2 score a sample of common CR items in Time 1, so that rater severity and discrimination in Time 2 can be put on

the same scale as Time 1. Tate's approach recognizes that, in the usual design, rater effects are confounded with examinee differences in the common-item non-equivalent (CINE) groups design. The IRT trend scoring approach allows one to disentangle the effects.

However, another approach to separating rater effects from item and examinee effects is to use a hierarchical rater model (HRM; Patz, 1996; Patz, Junker, Johnson, & Mariano, 2002), or its variation as the HRM signal detection theory (HRM-SDT) model (DeCarlo, 2010; DeCarlo, Kim, & Johnson, 2011), which is used here. The HRM-SDT model has important implications for the CINE design because it follows from the model that one does not need to do an additional linking study to adjust for rater differences over time. In particular, because the HRM-SDT model separates rater effects (Level 1) from item effects (Level 2), it provides CR item parameters that are not affected by rater severity and discrimination, but only by changes in the CR item parameters or the examinee distribution. Using simulations, Kim, DeCarlo, and Lee (2011) recently showed that when rater severity changed across Time 1 and 2 in the CINE design, along with examinee ability, the HRM correctly separated the effects and detected the simultaneous change in rater severity and examinee ability, whereas the IRT approach did not.

The present study further examines this issue and applies the HRM-SDT model to data from the Advanced Placement® (AP®) Studio Art portfolios 3-D Design, which operationally includes a trend scoring procedure to adjust for rater changes over time. The results for the HRM-SDT model are compared to results for the IRT trend scoring procedure. It is important to show that the HRM-SDT allows one to bypass the need for trend scoring for several reasons. First, trend scoring involves economic concerns – time and cost. Second, in order to keep time and effort at a reasonable level, a much smaller sample size is used for linking studies as compared to operational scoring. This means that the parameter estimates that are used for linking are of considerably poorer quality (e.g., larger standard errors) because of the small sample size. In contrast, the HRM approach to linking can be performed with the full operational datasets, given that a separate linking study is not needed, and so much more precise parameter estimates are obtained.

Rater Linking Methods for CR Items: IRT versus HRM-SDT approach

Studies have found that applications of traditional linking methods for CR items can lead to inaccurate results because changes in rater scoring over time can be confounded with changes in examinee ability. To resolve this problem, Tate (1999, 2000) suggested the IRT trend scoring method as discussed before. Tate (2003) and Kamata and Tate (2005) expanded the rater linking method to a long-term equating method for multiple years and conducted simulation studies to show the effectiveness of this method for mixed-format tests. Within the classical equating framework, Kim, Walker, and McHale (2010a, 2010b) examined the effectiveness of incorporating the rater linking method (they referred it to as *trend scoring*) in an equating design for mixed-format tests and CR tests in large-scale assessments. They showed that the common item equating design without rater linking produced biased results and that equating bias caused by rater severity change could be controlled by using a rater linking method.

Under an IRT model for CR items, the change in examinee ability between Time 1 and 2 can be represented by the following linear relationship, which involves using linking coefficients, the slope A and intercept B:

$$\theta^{(T1)} = A\theta^{(T2)} + B, \tag{1}$$

where $\theta^{(T1)}$ is the examinee ability on Time 1 scale and $\theta^{(T2)}$ the examinee ability on Time 2 scale. Given the relationship in Equation (1), the CR item parameters on the Time 1 and 2 scale can also be expressed using the linking coefficients.

For the generalized partial credit (GPC; Muraki, 1992) model, the item parameters, a_l (the item discrimination parameter) and b_{lm} (the item step parameter for category m of item l) on the Time 1 and Time 2 scale are related as follows:

$$\begin{aligned} a_l^{(T1)} &= a_l^{(T2)}/A \\ b_{lk}^{(T1)} &= Ab_{lk}^{(T2)} + B. \end{aligned} \tag{2}$$

When CR items are graded by different groups of raters across time, Tate (1999) pointed out that this relationship holds only when rater severity and discrimination are constant over time. Thus, Equation (2) can be re-written as:

$$\begin{aligned} a_l^{(T1,\,2)} &= a_l^{(T2,\,2)}/A \\ b_{lk}^{(T1,\,2)} &= Ab_{lk}^{(T2,\,2)} + B, \end{aligned} \tag{3}$$

where $a_l^{(T,\,J)}$ and $b_{lk}^{(T,\,J)}$ are the item parameters based on ratings from rating team J for Time T. Because different raters are usually used for Time 1 and Time 2, item parameter estimates for Time 1 associated with Time 2 raters, i.e., $a_l^{(T1,\,2)}$ and $b_{lk}^{(T1,\,2)}$ cannot be obtained using item calibration with data from Time 2. In order to find the Time 1 item parameter estimates for Time 2 raters, Tate (1999) suggested an additional linking study using trend scoring data where some of the raters at Time 2 also grade a sample of Time 1 examinee responses (to common CR items). Using the parameter estimates from the trend scoring data, the linking coefficients, which adjust for rater change over time, can then be obtained.

Although the IRT trend scoring procedure isolates any possible changes in rater severity or discrimination across Time 1 and 2, it requires storing Time 1 examinee responses to common CR items and then assigning them to some of the raters at Time 2. However, this trend scoring procedure is not necessary for the HRM-SDT model. Since the HRM-SDT model separates rater effects (Level 1) from item effects (Level 2), it provides item parameters that are adjusted for rater severity and discrimination, even if there are changes in the CR item parameters or the examinee distribution. Therefore, when the HRM-SDT model is used, the linking coefficients for CR items in Equation (2) can be directly obtained without an additional rater linking study.

The first level of the HRM-SDT model is a latent class signal detection model (DeCarlo, 2002; 2005; 2008; 2010). The SDT model models rater effects and it can be written as,

$$p(Y_{jl} \leq k \,|\eta_l = \eta) = F(c_{jkl} - d_{jl}\,\eta), \tag{4}$$

where Y_{jl} is the response of jth rater to the lth item, with the response being a discrete score k with K categories; η_l is a latent categorical variable for the lth item that takes on M values of η from 0 to $M-1$; and F is a cumulative distribution function (CDF). The parameter d_{jl} provides a measure of rater j's ability to discriminate between latent classes for the lth CR item. The second parameter, c_{jkl}, reflects the rater's use of response criteria and can indicate a variety of *rater effects*, such as severity or leniency, central tendency, and other effects (see Myford & Wolf, 2009).

The second level of the HRM-SDT treats the latent classes for each item (e.g., the latent essay categories) as ordinal indicators of examinee ability θ, using an IRT model such as the GPC model. Using GPC model, the second level of the HRM-SDT model can be written as

$$\log\left[\frac{p(\eta_l = \eta + 1|\theta)}{p(\eta_l = \eta|\theta)}\right] = a_l(\theta - b_{lm}), \tag{5}$$

where η_l is a latent categorical variable for item l that takes on values η from 0 to $M-1$ (it is assumed here that the number of latent classes, M, is the same as the number of response categories given in the scoring rubric, K, but this is not required); θ is a latent continuous variable (examinee ability) assumed to be N(0,1); a_l is an item discrimination parameter for the lth CR item; and b_{lm} are $M-1$ category step parameters, with $m = \eta+1$ (so that the step parameters are b_{l1}, b_{l2}, and so on). The HRM-SDT model simultaneously estimates rater parameters (d and c) and item parameters (a and b). The item parameters in Level 2 are adjusted for rater effects. Further details about the model are provided in DeCarlo (2010) and DeCarlo, Kim, and Johnson (2011).

METHOD

Data

The present study uses data from two administrations (2010 and 2011) of the AP Studio Art portfolios 3-D Design. AP Studio Art exams are designed for students who are seriously interested in the practical experience of art. The submitted portfolios by students are reviewed by raters consisting of college, university, and secondary school art instructors. Each of the portfolios consists of three sections, which can be viewed as three items. For the 3D exam, the sections are as follows: the Quality section (Section I) asks students to select five works that best exhibit a synthesis of form, technique, and content; the Concentration section (Section II) asks students to submit twelve images that demonstrate a depth of investigation and process of discovery; the Breadth section (Section III) asks students to submit eight works that demonstrate a serious grounding in visual principles and material techniques. The Quality section is graded by three raters while the Concentration and Breadth sections are graded by two raters. Students' weighted composite scores on these three sections are then converted to a grade on a 5 point scale. The AP grade qualification definitions are: 5—Extremely well qualified, 4—Well qualified, 3—Qualified, 2— Possibly qualified, and 1—No recommendation. AP exam grades of

5 are considered equivalent to a grade of A in corresponding college courses; AP exam grades of 4 are equivalent to grades of A−, B+, and B; AP exam grades of 3 are equivalent to grades of B−, C+, and C.

The 3D-Design exam sample contains 3,390 examinees from the 2011 administration and 3,178 examinees from the 2010 administration. For each administration, more than 100 raters were assigned to grade students' portfolios. In the current study, the scores from multiple raters on each item were collapsed and were treated as one score. That is, there were three scores for each quality rating, for example, however the scores came from different raters across different examinees. The scores were collapsed across raters, and so there were simply three scores for each examinee for quality, for example. The focus of the current study is on overall year to year change in rater scoring. Thus, there were three ratings for the Quality section (collapsed across raters) and two ratings for the Concentration and Breadth sections of the AP exam. The sample examined here contains trend scoring data for 126 examinees from the 2010 administration, which were randomly selected to represent the score distribution of the 2010 administration.

Note that the usual type of CR item, say an essay, has to be changed regularly because it is easily memorized and so security is compromised (McClellan, 2010). However, test security risks from using the same items over time are not of concern for AP Studio Art exams. Items in the Studio Art exams evaluate various aspects of art portfolios that students submit, and so knowing what the item is (i.e., "provide artworks that indicate the quality of your work") does not confer any advantage to examinees. Thus, the item questions for the Studio Arts exams do not have to be changed and the design is a common-item non-equivalent group design.

Two basic aspects are considered in the analysis: the first involves using the linking data to discover rater effects and the second involves using the linking data to get linking coefficients that describe examinee differences over time. The results from the IRT trend scoring approach are compared to the ones from the HRM-SDT approach solely with the operational data sets.

RESULTS

Rater Differences

Table 1 shows the mean scores across both the linking data and the full data. The top part of the table shows results for 126 cases from the linking dataset. This is of particular interest because it has the same examinees but different raters, namely the 2010 and 2011 set of raters, who both rated the same 2010 items. Because the examinees are the same across the two sets of data, any differences must be due to the different set of raters. Table 1 shows that the mean scores for Quality were slightly lower in 2011. For Concentration and Breadth, the mean scores were slightly higher in 2011. Thus, raters seemed to be slightly more severe in 2011 for Quality, but more lenient for Concentration and Breadth.

TABLE 1
Descriptive Statistics for AP Studio Art 3-D Design Data

	Quality		Concentration		Breadth	
	2010	2011	2010	2011	2010	2011
			Trend Scoring Data			
N	126	126	126	126	126	126
Mean	10.48	10.33	6.38	6.58	5.98	6.24
Std Dev	3.28	3.25	2.5	2.3	2.01	2.25
Minimum	3	3	2	2	2	2
Maximum	18	17	12	12	10	12
			Full Data			
N	3178	3390	3178	3390	3178	3390
Mean	10.29	10.32	6.41	6.55	6.17	6.36
Std Dev	3.25	3.24	2.48	2.41	2.21	2.34
Minimum	3	3	2	2	2	2
Maximum	18	18	12	12	12	12

Table 2 shows the parameter estimates for severity (*b*) and discrimination (*a*) for the 2010 operational data and the trend scoring data, which were obtained from the IRT trend scoring approach. With respect to rater "severity." the table shows that the average *b*'s in the trend scoring data were slightly lower in 2010 for Quality (mean of −0.06), lower for Concentration (−0.23), and higher for Breadth (0.10). This suggests greater leniency in 2011 for Quality and Concentration, but greater severity for Breadth. These results conflicted with the means discussed in Table 1.

TABLE 2

GPC Item Parameter Estimates

		2011		2010		Trend		Rater Difference
		Estimate	SE	Estimate	SE	Estimate	SE	Trend – 2010
Quality	b_1	–5.1236	0.3147	–5.9473	0.4085	–5.2099	1.7070	0.7374
	b_1	–2.1671	0.1428	–2.5319	0.1839	–2.2608	0.8197	0.2711
	b_1	0.2918	0.0684	0.3323	0.0834	–0.0345	0.3335	–0.3668
	b_1	2.4261	0.1606	2.8496	0.2112	2.4088	0.8285	–0.4408
	b_1	4.7832	0.2978	5.9301	0.4113	5.4434	1.6797	–0.4867
	a_1	2.2136	0.1569	2.7220	0.2145	2.1377	0.8039	–0.5843
Concentration	b_2	–2.8592	0.1398	–2.8178	0.1340	–3.3040	0.8941	–0.4862
	b_2	–1.2526	0.0727	–0.9412	0.0690	–1.6852	0.4205	–0.7440
	b_2	0.1579	0.0526	0.2394	0.0553	0.3228	0.2626	0.0834
	b_2	1.327	0.0767	1.2268	0.0788	1.7039	0.4579	0.4771
	b_2	2.0447	0.1223	2.0509	0.1243	1.5662	0.6736	–0.4847
	a_2	1.1684	0.0598	1.1207	0.0591	1.1612	0.3442	0.0405
Breadth	b_3	–2.6465	0.1239	–2.7722	0.1282	–2.0405	0.5457	0.7317
	b_3	–1.0859	0.0638	–1.0250	0.0629	–0.9245	0.3113	0.1005
	b_3	0.3134	0.0510	0.4576	0.0527	0.1533	0.2517	–0.3043
	b_3	1.2661	0.0738	1.4054	0.0799	1.4805	0.3805	0.0751
	b_3	2.1638	0.1200	2.6534	0.143	2.5605	0.7294	–0.0929
	a_3	1.0105	0.0480	1.0061	0.0485	0.8517	0.2171	–0.1544

The HRM-SDT model approach, on the other hand, followed from the point made in Kim et al. (2011) that one can simply use the operational datasets, and not the trend data, and make valid comparisons with respect to examinee and rater differences, given that they are separated by the HRM-SDT model. Table 3 shows the relative criteria locations from the HRM-SDT model, which take into account rater magnitude of discrimination parameter and thus allow the direct comparisons of rater severity between rater response criteria parameters from the 2011 and 2010 operational datasets. The average difference in the criteria locations between 2011 and 2010 data for Quality was small and positive (0.003) and for both Concentration and Breadth were negative (−0.025 and −0.049). Thus, the HRM-SDT model indicated that the raters were about the same, perhaps slightly more severe for quality in 2011. For Concentration and Breadth, the raters were more lenient in 2011. Both of these results were perfectly consistent with the analysis of the means discussed above.

In sum, the IRT analysis using the linking data gave somewhat inconsistent results with respect to the analysis of the mean scores in the linking data whereas the HRM results were consistent with respect to the mean score analysis.

TABLE 3
HRM-SDT Relative Criteria Locations

		2011	2010	Rater Difference (2011 – 2010)
Quality	c_{111}	0.0006	–0.0243	0.0249
	c_{112}	0.2973	0.2838	0.0136
	c_{113}	0.5551	0.564	–0.0092
	c_{114}	0.7922	0.8179	–0.0256
	c_{115}	1.0266	1.0590	–0.0324
	c_{121}	0.0062	–0.0219	0.0281
	c_{122}	0.2868	0.2745	0.0124
	c_{123}	0.5601	0.563	–0.0029
	c_{124}	0.8202	0.8222	–0.0019
	c_{125}	1.0450	1.0489	–0.0039
	c_{131}	0.0073	–0.0349	0.0422
	c_{132}	0.2993	0.2564	0.0429
	c_{133}	0.5656	0.5677	–0.0022
	c_{134}	0.8048	0.8200	–0.0152
	c_{135}	1.0281	1.0602	–0.0321
Concentration	c_{211}	0.0518	0.0703	–0.0185
	c_{212}	0.3012	0.3473	–0.0462
	c_{213}	0.5464	0.5953	–0.0489
	c_{214}	0.7714	0.8028	–0.0314
	c_{215}	1.0088	1.0016	0.0072
	c_{221}	0.0876	0.0717	0.0159
	c_{222}	0.3276	0.3368	–0.0092
	c_{223}	0.5717	0.6005	–0.0288
	c_{224}	0.7747	0.8280	–0.0533
	c_{225}	0.9967	1.0370	–0.0403
Breadth	c_{311}	0.0490	0.0493	–0.0002
	c_{312}	0.3500	0.3716	–0.0217
	c_{313}	0.6114	0.6678	–0.0565
	c_{314}	0.8317	0.9095	–0.0778
	c_{315}	1.0507	1.1526	–0.1019
	c_{321}	0.0498	0.0359	0.0139
	c_{322}	0.3417	0.3629	–0.0212
	c_{323}	0.6017	0.6544	–0.0527
	c_{324}	0.8266	0.9025	–0.0759
	c_{325}	1.0422	1.1381	–0.0959

Examinee Differences

The linking coefficients with the IRT trend scoring approach and HRM-SDT model approach were obtained using four methods – Mean/Mean, Mean/Sigma, Stocking-Lord (Stocking & Lord, 1983), and Haebara (1980) – through POLYST (Kim & Kolen, 2003). Table 4 presents results for the linking coefficients obtained from both IRT trend scoring approach and HRM-SDT model approach. The table shows that, for both approaches, A is close to 1.0 and B is close to zero. Given that both approaches give the similar conclusions, this shows that one might be able to bypass the time and expense involved with obtaining trend scoring data.

TABLE 4

Linking Coefficients Obtained for GPC and HRM

	GPC		HRM	
Method	Slope	Intercept	Slope	Intercept
Mean/Mean	1.0583	0.0375	0.9703	0.0036
Mean/Sigma	1.0574	0.0374	0.9402	-0.0001
Haebara	1.0230	-0.0760	0.9015	-0.0314
Stocking-Lord	1.0401	-0.0103	0.9191	-0.0516

PRATICAL AND POLICY IMPLICATIONS

The present study demonstrates an extremely important practical aspect of the HRM-SDT, namely that one does not need to include a trend scoring procedure to adjust for rater differences over time. This can be avoided with the HRM-SDT because it separates rater effects from item and examinee effects, whereas these are confounded in the IRT approach, which is why the latter requires a linking study. The importance of this cannot be overemphasized, given that trend scoring requires considerable time and money. Given the widespread use of CR items in many assessments, further studies of the HRM-SDT and its role for rater-linking are needed.

REFERENCES

DeCarlo, L. T. (2002). A latent class extension of signal detection theory, with applications. *Multivariate Behavioral Research*, *37*,423-451.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement*,*42*, 53-76.

DeCarlo, L. T. (2008). *Studies of a latent-class signal-detection model for constructed response scoring* (ETS Research Report. No. RR-08-63). Princeton NJ: ETS.

DeCarlo, L. T. (2010). *Studies of a latent class signal detection model for constructed response scoring II: Incomplete and hierarchical designs* (ETS Research Report No. RR–10–08). Princeton NJ: ETS.

DeCarlo, L. T., Kim, Y. K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*, 333–356.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares methods. *Japanese Psychological Research, 22*, 144–149.

Kamata, A., & Tate, R. L. (2005). The performance of a method for the long–term equating of mixed–format assessment. *Journal of Educational Measurement, 42,* 193–213.

Kim, Y. K., DeCarlo, L. T. & Lee, W. (2011). *On implications of a Hierarchical Rater/Signal Detection Model for IRT linking with constructed response items.* Paper presented at the 2011 meeting of the National Council on Measurement in Education, New Orleans, LA.

Kim, S., & Kolen, M. J. (2003). *POLYST: A computer program for polytomous IRT scale transformation* [Computer software]. Retrieved May 1, 2010, from http://www.education.uiowa.edu/casma/

Kim, S., Walker, M. E., & McHale, F. (2010a). Comparisons among designs for equating mixed–format tests in large–scale assessments. *Journal of Educational Measurement, 47,* 36–53.

Kim, S., Walker, M. E., & McHale, F. (2010b). Investigating the effectiveness of equating designs for constructed–response tests in large–scale assessments. *Journal of Educational Measurement, 47,* 186–201.

McClellan, C. A. (2010, February). Constructed response scoring — doing it right. *R & D Connections, 13*, 1 – 7.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied* Psychological Measurement, 16, 159–176.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, 46(4), 371-389.

Patz, R. J. (1996). Markov Chain Monte Carlo methods for item response theory models with applications for NAEP. Ph.D. dissertation, Carnegie Mellon University, United States — Pennsylvania. Retrieved September 14, 2008, from Dissertations & Theses: Full Text database. (Publication No. AAT 9713184).

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large−scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341−384.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Tate, R. L. (1999). A cautionary note on IRT–based linking of test with polytomous items. *Journal of Educational Measurement, 36*, 336−346.

Tate, R. L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*, 329−346.

Tate, R. L. (2003). Equating for long–term scale maintenance of mixed format tests containing multiple choice and constructed response items. *Educational and Psychological Measurement*, *63*, 893–914.

KAERA Research Forum, *1(1)*, 36-46

RESEARCH ARTICLE

The College Scholastic Ability Test in Korea: Introduction and Related Issues[1]

Chanho Park, Ph.D.
Keimyung University

After nine years of compulsory education, most Korean students take academic route at high schools for higher education, and the rate of graduation from high schools is high. The Korean government maintains control over the college entrance examination, which once belonged to colleges and universities. The college scholastic ability test (CSAT) was developed as a norm-referenced test in a paper-and-pencil format to measure higher-level thinking abilities. It is the only national college entrance examination currently used and thus plays an important role in the education system in Korea. Korea Institute for Curriculum and Evaluation (KICE) is commissioned by the government to take charge of the CSAT. The CSAT has five domains, and in some domains examinees can choose up to three subjects as electives. The numbers of items are between 20 and 50, and except for nine Mathematics short-answer items, all remaining items are in multiple-choice format. KICE conducts psychometric analyses on the CSAT using both classical test theory and item response theory, and the results are acceptable. However, CSAT's power of predicting college grade point averages has been found poorer than that of high school records (Kim, 2010). Now the CSAT reports bounded standard scores, percentiles, and stanines. The CSAT receives national attention and thus requires strong measures to secure the test forms. Although the CSAT is still an important factor for college admissions, as universities start to increase weights on other factors such as high school records, some changes are planned on the CSAT, and more changes are being discussed.

Keywords: College Scholastic Ability Test, College Admissions, College Entrance Policies, Test Security

Before entering colleges or universities, Korean students attend elementary schools for six years, middle schools for three years, and then high schools for three years. Education is compulsory from elementary to middle school (from ages 7 to 15), and although high school education is not compulsory, most middle school graduates go to high schools. There are many types of high schools, but the majority of the schools are general schools where students take academic route. The number of high schools whose graduates took the college entrance examination was 2,297 in

[1] This paper was presented at the First International Conference for Assessment& Evaluation, December 2-4, Riyadh, Saudi Arabia. Note that this paper does not reflect the latest changes in the college scholastic ability test in Korea.

the year 2011. Among them, about 66 percent (i.e., 1,520 schools) were general high schools, and the rest were vocational high schools, mixed type, etc.

Education is deemed important for Koreans. Most students complete secondary education with a graduation rate from secondary education of 89 percent (OECD, 2011). A high proportion of secondary school graduates also pursue higher education. The college scholastic ability test (CSAT), which is the only national college entrance examination currently used, is still an important factor for college admissions. The influence of the CSAT is not limited to the educational system of Korea. It somehow affects the whole society. Due to Korean's zeal for higher education, the examination and policies related to college admissions have played an important role in the educational system of Korea. Thus, although many budgetary and administrative decisions were delegated to municipal and provincial district offices of education, the central government maintains control over the college entrance examination.

In this paper, a brief history of the college entrance examination and policies in Korea is reviewed first. The domains and subjects of the CSAT, the item format, the number of items and examinees, and the test time are summarized. Then, test development procedures and psychometric properties of the CSAT are examined, which is followed by an investigation into how the test is used for college admissions. The paper concludes by discussing the social aspects of the CSAT and future changes caused by the considerations of social needs.

HISTORY OF COLLEGE ENTRANCE EXAMINATION AND POLICIES

After the Korean War, the colleges and universities in Korea administered their own screening tests until 1960s. However, having as many screening tests as the number of colleges or universities was costly, and fairness became a social issue. Thus, in the year of 1969, the Korean Ministry of Education (MOE) decided to administer the national preliminary examination to screen unqualified applicants before they take main examinations at the colleges or universities they applied to.

The national preliminary examination provided a common basic standard to college applicants, and high schools educated their students so that the students could successfully prepare for the test. It therefore helped standardize secondary education in Korea. However, the purpose of the preliminary examination was only to screen the applicants below the minimum acceptance level. As college admissions became more competitive as the "baby boomers" after the Korean War became of age to start higher education, the MOE decided to develop a valid and fair college entrance examination unifying the preliminary examination and main examinations, which can be used by all colleges and universities.

A new college entrance examination, the scholastic achievement test, started in the year 1982. The scholastic achievement test was a norm-referenced test based on high school curriculum. The colleges and universities could no longer have their own tests and had to recruit students mostly based on the new scholastic achievement test. Although high school grades played a role in college admissions, the results of the scholastic achievement test were the most significant factor. The test was used until 1992. While the test was used for about ten years, criticisms arose that students were studying through rote memorization to prepare for the test, and that high schools were becoming prep schools for the test. Also, private tutoring outside of schools became prevalent, which caused another concern, and the need for a new test was raised.

COLLEGE SCHOLASTIC ABILITY TEST

Overview

The CSAT is the only national college entrance examination authorized by the government. It is a norm-referenced test administered in paper-and-pencil format. It was launched in 1993 after four years of pilot testing. Introduction of the CSAT was motivated by the necessities for measuring higher-level thinking abilities. For that purpose, test materials were taken from outside the textbooks, while the scholastic achievement test, the predecessor of the CSAT, was exclusively based on textbooks. That is, the CSAT was developed to overcome the limitations of the scholastic achievement test, which could be prepared for by rote. The purpose of the CSAT was to screen qualified candidates for college education, to set a guideline for secondary education, and to provide colleges or universities with fair and objective data about the applicants.

The CSAT is a standardized test accepted by all colleges and universities in Korea. Except for nine short-answer items in Mathematics, all other test items are multiple choice items with five answer choices. By taking the CSAT, examinees receive a rank or score that can only be used in support of the current year's applications for universities. Students wishing to defer university entrance must retake the CSAT in the year they wish to enter university.

Testing Organization of the CSAT

The MOE of Korea commissions research, management, and administration of the CSAT to Korea Institute for Curriculum and Evaluation (KICE). Since its inception, KICE has undertaken the CSAT as one of its main projects.

KICE was established in 1998 as an educational research institute. It has been funded by the government to contribute to the improvement of elementary and secondary education as well as the nation's educational development through research, development, and implementation of curriculum and educational evaluation. The areas of research include curriculum and instruction, educational assessment, authorization of elementary and secondary school textbooks, etc. (Korea Institute for Curriculum and Evaluation, 2012). KICE is also commissioned to develop and administer other national tests such as the national assessment of educational achievement, national English ability test, etc., and is participating in international comparative studies such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA).

About 400 full-time employees work at KICE, among whom more than 150 researchers have doctoral degrees. The majors of the researchers cover almost all fields of study in education such as educational measurement or evaluation, educational psychology, curriculum and instruction, and education of each subject the national curriculum covers.

Structure of the CSAT

Although the principle and basic functions of the CSAT remain unchanged, there have been changes with regard to the domains and score reporting. The first version of the CSAT had 190

items on three domains: Korean Language, Mathematics, and English as a Foreign Language. The first version was characterized by the test items based on interdisciplinary materials on all domains. As for test results, only raw scores were reported with a range from 0 to 200.

Having undergone changes big and small, the CSAT now consists of five domains: Korean Language, Mathematics, English as a Foreign Language, Social Studies/Science/Vocational Education, and Foreign Languages/Chinese Characters and Classics. Examinees can freely choose all or some of the five domains (see Table 1). Among the five domains, only Korean Language and English as a Foreign Language use interdisciplinary materials now. For Mathematics, two types (A and B) are available. Students wishing to major in natural science or engineering usually take Math A, while those who want to major in humanities or social science select Math B.

For the domain of Social Studies/Science/Vocational Education, students first select one of three sub-domains, and within a sub-domain students can choose up to three subjects as electives. The eleven subjects for the Social Studies sub-domain are Ethics, Korean History, Korean Modern and Contemporary History, World History, Politics, Economics, Society and Culture, Law and Society, Korean Geography, Economic Geography, and World Geography. Eight Science subjects are Physics I and II, Chemistry I and II, Biology I and II, and Earth Science I and II. For Vocational Education, examinees can also choose up to three subjects, but only one is allowed out of the four subjects—Agricultural Information Management, Basic Information Technology, General Computers, and Fishery and Shipping Information Processing—and up to two can be selected from the following 13 subjects: Understanding of Agriculture, Techniques in Basic Agriculture, Introduction to Industry, Basic Drafting, Commercial Economy, Principles of Accounting, Introduction to Fisheries, General Marine Affairs, General Oceanography, Human Development, Food and Nutrition, General Design, and Programming. Examinees can select one out of the eight subjects—German I, French I, Spanish I, Chinese I, Japanese I, Russian I, Arabic I, and Chinese Characters and Classics—in the domain of Foreign Languages/Chinese Characters and Classics.

TABLE 1
Domains and Subjects

Domain		Subjects
Korean Language		N/A
Mathematics (Select 1)	Math A	
	Math B	
English as a Foreign Language		
Social Studies / Science / Vocational Education (Select 1)	Social Studies (Select up to 3)	Ethics, Korean History, Korean Modern and Contemporary History, World History, Politics, Economics, Society and Culture, Law and Society, Korean Geography, Economic Geography, and World Geography
	Science (Select up to 3)	Physics I & II, Chemistry I & II, Biology I & II, Earth Science I & II
	Vocational Education (Select up to 3)	One of the following: Agricultural Information Management, Basic Information Technology, General Computers, and Fishery and Shipping Information Processing Up to two of the following: Understanding of Agriculture, Techniques in Basic Agriculture, Introduction to Industry, Basic Drafting, Commercial Economy, Principles of Accounting, Introduction to Fisheries, General Marine Affairs, General Oceanography, Human Development, Food and Nutrition, General Design, and Programming
Foreign Languages / Chinese Characters and Classics (Select 1)		German I, French I, Spanish I, Chinese I, Japanese I, Russian I, Arabic I, and Chinese Characters and Classics

The number of items and the test time are shown in Table 2. The Korean Language domain has 50 multiple choice (MC) items including five listening comprehension items, and examinees are given 80 minutes. The 30 items on Mathematics should be answered within 100 minutes. In Mathematics, nine items (30%) have a short-answer format while the remaining 21 are MC items. English as a Foreign Language also has 50 items, and 17 of them belong to the listening comprehension section. 70 minutes are given for the 50 MC items. Each subject in the domain of Social Studies/Science/Vocational Education has 20 MC items, and 30 minutes are given per subject. 40 minutes are given to the selected subject in Foreign Languages/Chinese Characters and Classics, which has 30 MC items. All MC items have five answer choices.

TABLE 2

Number of Items, Test Time, and Item Format on the CSAT

Domain		Number of Items	Test Time (mins)	Item Format
Korean Language		50 (5 listening)	80	Multiple Choice
Mathematics (select 1)	Math A / Math B	30	100	Multiple Choice (70%) Short Answer (30%)
English as a Foreign Language		50 (17 listening)	70	Multiple Choice
Social Studies / Science / Vocational Education (select 1)	Social Studies / Science / Vocational Education	20 per subject (Up to 3 subjects)	Up to 90: 30 per subject	Multiple Choice
Foreign Languages / Chinese Characters and Classics		30	40	Multiple Choice

Examinees of the CSAT

The number of CSAT examinees is generally between 500,000 and 700,000 per year. In 2011, 648,946 examinees took the CSAT. Among them, 494,057 examinees were 12th-grade students taking the CSAT for the first time, and 141,211 examinees were high school graduates, most of whom were repeaters. Table 3 shows the numbers of examinees in the last three years. The number of examinees was the largest in the year of 2010, and is expected to gradually decrease from 2011 on.

TABLE 3

Number of Examinees per Type in the Last Three Years

Type	2009	2010	2011
High School Students	503,092	510,893	494,057
High School Graduates	121,877	144,056	141,211
Others (e.g., GED Passers)	13,247	14,042	13,678
Sum Total	638,216	668,991	648,946

Administration Procedures of the CSAT

The CSAT is administered once a year except for the first administration in 1993, when two test forms were developed and administered. The test date is generally the second Thursday of November. It takes almost a full day to take all domains or subjects. Table 4 shows the administration procedures of the CSAT. Starting from 8:40 AM, the test ends at 5:35 PM if examinees chose a subject in Foreign Languages/Chinese Characters and Classics. Time is allocated according to the test time shown in Table 2. However, four minutes are added to the domain of Social Studies/Science/Vocational Education since two minutes are needed after one elective is finished to prepare for the next elective.

TABLE 4
Administration Procedures of the CSAT

Session	Domain	Time
	Entering the room by 08:10	
1st	Korean Language	08:40 – 10:00 (80 mins)
	Break: 10:00 – 10:30 (30 mins)	
2nd	Mathematics	10:30 – 12:10 (100 mins)
	Lunch: 12:10 – 13:10 (60 mins)	
3rd	English as a Foreign Language	13:10 – 14:20 (70 mins)
	Break: 14:20 – 14:50 (30 mins)	
4th	Social Studies/Science/Vocational Education	14:50 – 16:24 (94 mins)
	Break: 16:24 – 16:55 (31 mins)	
5th	Foreign Languages/Chinese Characters and Classics	16:55 – 17:35 (40 mins)

Psychometric Properties of the CSAT

After each year's administration, KICE regularly conducts classical psychometric analyses. The analyses include point-biserial correlation coefficients for item discrimination, item endorsement rates for item difficulty, and Cronbach's alpha for test reliability (Crocker & Algina, 1986). The results of these analyses are for internal use only and are not published. Some of the results of psychometric analyses for the 2011 data are as follows: The alpha coefficients are up to .95 in the domains of Korean Language, Mathematics, and English as a Foreign Language, and are about .90 for the 20 or 30 items in the subjects of the other domains. Average point-biserial correlation coefficients per subject are between .3 and .6. The average item endorsement rate is about .6, as the test blueprint specifies.

Special studies are conducted occasionally. For example, an internal research study was conducted to investigate the comparability between standard scores (a linear transformation of raw scores) and latent trait scores obtained under the framework of item response theory (IRT). Although standard scores are being used for the CSAT, IRT-based scores have continuously been studied. Research also includes methods of multiple administrations per year and possible plans for test equating, but no particular decisions have been made yet.

Before applying IRT, dimensionality of the CSAT data was tested to examine the unidimensionality assumption of IRT. For dimensionality assessment, DIMTEST (Stout, 1987) and DETECT (Zhang & Stout, 1999) were applied as nonparametric analyses, and TESTFACT (Wood et al., 2003) was used as a parametric analysis. The results of these analyses all confirmed unidimensionality of the domains of Korean Language, Mathematics, and English as a Foreign Language. Dimensionality assessment was conducted only on these three domains.

Validity of the CSAT has been found acceptable. Since validating a test could mean gathering supportive evidence in many aspects of the test (AERA, APA & NCME, 1999), continuous efforts are made for validation of the CSAT. Although validity of the CSAT has been found to be acceptable for contents or response processes, the CSAT is often questioned on how well the results predict grade point averages (GPAs) in colleges or universities. High school grades were generally found to be better predictors of college GPAs, which motivated more universities to increase the weight of high school grades as a factor for admissions (Kim, 2010).

USE OF THE CSAT

Test Scores

When the CSAT was introduced in the year 1993, only raw scores (i.e., weighted sum of item secores) and percentiles were reported. In its first administration, two forms of the CSAT were administered in order to reduce the anxiety of the students on such a high-stakes test. However, since all test items are released to the public right after the administration, test equating (i.e., statistical moderation of test scores for comparability of different forms of a test; Kolen & Brennan, 2004) could not be implemented; thus, the same students obtained different raw scores mainly due to the different difficulty levels of the two forms. This caused confusions, and since then, the CSAT has been administered only once a year.

Although the CSAT is valid only for the year's admission, it was not easy for colleges and universities to interpret raw scores, which depend on the difficulty level of the test as well as examinees' abilities. In addition, electives were introduced in the year 1998, and students could choose among the subjects in some domains. Standard scores (T-score and its variants) thus began to be used among the electives within a domain. Now, the CSAT reports standard scores, stanines, and percentiles on all domains, and raw scores are no longer reported.

The scoring process is as follows (see Table 5 for a summary). First, raw scores are calculated for each domain or subject as weighted sums of item scores. The maximum raw scores are 100 for Korean Language, Mathematics, and English as a Foreign Language and 50 for electives in the other domains. T-scores (i.e., standard scores with a mean of 50 and standard deviation of 10) are used for the domains of Social Studies/Science/Vocational Education and Foreign Languages/Chinese Characters and Classics. In these domains, standard scores are truncated by the lower and upper bounds of 0 and 100, respectively. For the domains of Korean Language, Mathematics, and English as a Foreign Language, standard scores have a mean of 100 and standard deviation of 20 with a lower bound of 0 and an upper bound of 200.

Once standard scores are obtained, they are rounded to integers, and percentiles are calculated based on the rounded standard scores. Grades are also reported as reversed stanine grades; that is, the top 4 percent of the examinees receive a grade of 1, the next 7, 12, 17, 20, 17, 12, 7, and 4 percent receive grades of 2 to 9, respectively.

TABLE 5
Standard Scores for the CSAT

Domain	Number of Items	Maximum Raw Score	Standard Score		
			Mean	Standard	Range*
Korean Language	50	100	100	20	0 –200
Mathematics	30				
English as a Foreign Language	50				
Social Studies / Science / Vocational Education	20	50	50	10	0 –100
Foreign Languages / Chinese Characters and Classics	30				

* Range refers to the lower and upper bound for truncation.

CSAT for College Admissions

12th-grade students and high school graduates or their equivalents apply directly to the college or university they wish to attend (Ministry of Education, Science and Technology, 2012). When the national preliminary examination was used only to screen qualified candidates, colleges and universities recruited students mostly based on the main examinations of their own. When the main examinations were prohibited, the first criterion universities used was the scores on the scholastic achievement test. Although high school records were also considered, they only had minor influences on admissions.

When the CSAT was first administered, its influence on admissions was similar to that of the scholastic achievement test. Most universities recruited students by regular admissions, for which CSAT results were the most important factor. It was difficult for examinees with good high school records to enter the university they desired if their CSAT scores were low. Although the CSAT as a valid and reliable assessment tool provided the colleges and universities with objective data about the examinees, admissions based on the results of only one test could be a problem.

From the late 1990s, colleges and universities began to consider other factors for admissions. These days, universities judge applicants based on a variety of factors such as student's high school records including continuous teacher assessment, CSAT results, involvement in extracurricular activities, teacher recommendation, student's essays, etc. As more factors are being considered, methods other than regular admissions are gaining popularity. The influences of the CSAT are gradually decreasing, and more universities are now employing assessment specialists as admission officers, who can make a professional judgment on students' qualifications based on diverse factors. Sometimes, admissions are made even without CSAT scores.

Test Security and Social Dimensions

Although the influences of the CSAT on college admissions are decreasing, many people still regard the CSAT as the most important factor for admissions. Thus, test security is also an important issue. To minimize the risk of security breach, items for the CSAT start to be developed a month before the test administration. Item writers have to stay in the "item writing camp" while the test forms are developed and printed. The camp is completely blocked from the outside, and no contact is allowed while staying in the camp for 32 days until the test date. On the test date, examination papers are carried to the test sites under the protection of the police. Thus, test items are almost perfectly secured until the moment examinees receive them.

Inside the test room, examinees are strictly prohibited from carrying any electronic devices. Two proctors per test room monitor the examinees. If cheating is detected, the examinee cannot retake the CSAT for a period of time depending on the severity of the cheating. Cheating is only detected on the site, and no further analyses are conducted with regard to cheating once the test is finished.

The CSAT receives national attention. While examinees take listening comprehension sections of Korean Language and English as a Foreign Language, airplanes are not allowed to land or take off so that examinees nearby are not disturbed by unpredicted noises. The CSAT is also influential to secondary school education. Many high schools want to spend as much time as possible on the domains of Korean Language, Mathematics, and English as a Foreign Language because these domains are more influential than the others. Also, schools try to spend as little time as possible on the subjects that are not covered by the CSAT such as music and physical education.

Since the CSAT was developed to measure higher-level thinking abilities, one of its characteristics is the use of interdisciplinary materials. In order to prevent students from simply memorizing textbooks, test materials of the CSAT were taken from outside of the textbooks. For example, some reading passages in the Korean Language domain are taken from natural science or engineering texts. While this attempt was welcomed by universities, it caused other problems. Since most classes in high schools used only textbooks as teaching materials under the national curriculum system, students began to search for the places where they can learn about those test materials, and extracurricular private schools or tutors specialized for CSAT preparation began to flourish.

The Korean MOE has tried to reduce private tutoring and to restore the functions of schools. One of the measures the MOE took was to adopt test materials from the contents of Educational Broadcasting System, a government-funded broadcasting system for public education in Korea. Also, in order to reduce the burden of private tutoring for students and their families, the MOE has been trying other measures, such as maintaining the difficulty of the CSAT at an easy level and decreasing the maximum number of electives. In addition, from the year 2013, the number of the subjects in Social Studies/Science/Vocational Education is reduced, while Vietnamese is added in the domain of Foreign Languages/Chinese Characters and Classics. Also, the domains of Korean Language and English as a Foreign Language are divided into levels A and B like Mathematics, and test materials of Korean Language and English as a Foreign Language are based on high school curriculum instead of interdisciplinary materials.

DISCUSSION AND CONCLUSION

As the Korean society underwent substantial changes after the Korean War, the college entrance examination and policies also experienced many changes. Colleges and universities once selected students based on their own tests, and then the national college entrance examination was introduced, which helped standardize secondary education. Education in Korea adopts the national curriculum, to which the national college entrance examination has been closely related. However, the CSAT's use of test materials outside the textbooks and its high difficulty levels could undermine the national curriculum system as students resort more to private prep schools to prepare for the test.

The changes in the CSAT from the year 2013 are made with an aim to restore the circulatory relationship between the CSAT and the national curriculum. In addition, both the number of subjects and the maximum number of electives decrease, and both Korean Language and English as a Foreign Language are divided into levels A and B so that students can choose a form that meets their levels. All of these changes are related to decreasing influences of the CSAT on college admissions.

More issues remain with regard to the future CSAT beyond the changes in 2013. In addition to the aforementioned issues of applying IRT and multiple administrations with test equating, one of the topics currently being discussed is whether to keep the CSAT as a norm-referenced test or to switch it to a criterion-referenced test, reporting only grades as qualifiers. Another topic with regard to the test format is the use of constructed-response (CR) items and essays. More complex skills will be needed in the 21st century, and CR items can be better suited for such complex skills than MC items. KICE keeps conducting research on these issues, but no conclusions have been made yet.

REFERENCES

AERA, APA, & NCME (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

Kim, S.-H. (2010). A diagnosis of the CSAT applying a logical model of validation. *Journal of Educational Evaluation*, *23*, 1-27.

Korea Institute for Curriculum and Evaluation (2012). *About KICE*. Retrieved May 5, 2012 from www.kice.re.kr/en/introduction/about.jsp.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Ministry of Education, Science and Technology (2012). *Higher education*. Retrieved October 3, 2012 from http://english.mest.go.kr/web/1697/site/contents/en/en_0207.jsp.

OECD (2011). *Education at a glance 2011: OECD indicators*. Paris, France: OECD.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617.

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. D. (2003). *TESTFACT: Test scoring and full information item factor analysis* (Version 4.0) [Computer software]. Lincolnwood, IL: Scientific Software International.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure, *Psychometrika*, *64*, 213-249.

KAERA Research Forum, *1(1)*, 47-54

RESEARCH ARTICLE

Analyzing Standard Setting Data Using a Generalizability Theory Framework

MinJeong Shin

University of Massachusetts Amherst

Standard setting has been widely used to determine levels of proficiency and cutscores corresponding to performance levels on a test. Standards are usually established by panelists' judgments and could be affected by various possible sources of error. This study applied a generalizability theory framework to Massachusetts Adult Proficiency Tests (MAPT) standard setting studies to investigate sources of error variance in a generalizability study (G-study) as well as the effect of the sample size for items or panelists in a decision study (D-study). Two different study designs were used: the item descriptor matching (IDM) method with the p x (i : f) design and the modified Angoff method with the (p : g) x i design. The results suggested that using more items was more effective in reducing absolute error variance than recruiting more panelists or developing more parallel test forms. Moreover, the results of applying the (p : g) x i design suggested that doubling the number of groups decreased error more effectively when the total number of panelists remained unchanged. Considering both the cost and time, sample sizes including panels, items, forms, and groups should be carefully determined within the range to minimize error.

Keywords: standard setting, generalizability theory

Standard setting procedures have been widely used to determine performance standards on assessments as demands for accountability in education have increased. Performance standards are established on the basis of judgments which can incorporate different sources of error including panelists or items. Although judgments cannot be perfectly objective, sufficient validity evidence is required to support classification decisions. According to Kane (2001), large discrepancies in the standard setting process may indicate a lack of internal validity. As a criterion of internal evidence can be the standard error of the cut score (Sireci et al., 2007), this study applied a generalizability theory (G-theory) framework to examine how discrepant panelists' judgments are and which component contributes most to the total error. The objectives of this study are to examine sources of error variance using a generalizability study (G-study) and to investigate the effect of changing the sample size of each facet such as items or panelists within a decision study (D-study).

Using different methods such as the Bookmark, Angoff, and Nedelsky methods, several studies applied the G-theory framework to examine the sources of variability in setting

performance standards (Brennan, 1995; Brennan & Lockwood, 1980; Chang, 1999; Chang & Hocevar, 2000; Clauser et al., 2009; Clauser, Margolis, & Clauser, 2012; Lee & Lewis, 2008; Kane & Wilson, 1984; Raymond & Reid, 2001; Yin & Sconing, 2008). Yin and Sconing (2008) as well as Lee and Lewis (2008) estimated the standard error of cutscores under the G-theory framework in various facets or conditions including the number of panels, items, test forms, and groups.

As G-theory makes it possible to differentiate multiple sources of error, applying it to standard setting studies allows researchers to analyze multiple factors affecting variability. In general, five possible facets can be considered in standard setting procedures: panelists, rounds, items, groups of panelists, and test forms. First, choosing panelists is an important task since the cutscores are based on their judgments. Panelists often share opinions and change their responses after a group discussion. Second, standard setting procedures usually span several rounds. Although the initial decision about setting cutscores may change over rounds, it has been observed that participants' judgments have less variation as the number of rounds increases. Third, selecting items and test forms representing an item pool is also crucial in standard setting.

In G-theory analysis, G- and D-studies are conducted for different purposes. The G-study is used to provide estimates of the variability of possible facets of measurement. That is, based on this analysis, researchers are able to identify the magnitude of each source of error and estimate the error variance. The D-study focuses on the specification of a universe of generalization. Using the information obtained from the G-study, a decision-maker can choose conditions that s/he wants to generalize.

The current study applied the G-theory framework to the Massachusetts Adult Proficiency Tests (MAPT), which employed two standard setting methods – the item-descriptor matching (IDM) method and the modified Angoff method. As its name implies, the IDM method requires panelists to match item responses to the description of performance in terms of the knowledge and skills expected of examinees in a certain performance level. What is required of panelists is to identify the response that is closely aligned with the description of the proficiency category. The Angoff method involves judgments about performance of borderline learners who just exceed the threshold. In general, the Angoff method is based on predicting whether a borderline examinee would answer an item correctly, and on estimating the probability of getting an item correct.

METHOD

Data

Data were obtained from two MAPT standard setting studies for math and reading conducted in 2006 and one study for reading in 2007.The IDM method was employed for both math and reading tests in 2006, and the modified Angoff method was applied to re-evaluate reading cutscores in 2007. Because the goal of standard setting process for the MAPT was to identify cutscores corresponding to the National Reporting System's (NRS) Educational Functioning Levels (EFLs), applying a method focusing on examinees' knowledge and skills within each proficiency level was more useful than a method utilizing probability judgments based on the examinees' performance at the borderline of categories (Sireci et al., 2007).

In 2006, 10 and 11 panelists, who were either classroom teachers or ABE program administrators with prior classroom teaching experience, participated in standard setting for math and reading, respectively. A different group of twelve panelists participated in 2007. The training procedures were somewhat different depending on the standard setting method, either the IDM method or the modified Angoff method; however, the performance categories used in standard setting remained the same based on the EFLs. Details of these panelists and procedures are provided in Sireci et al. (2008). For the IDM method, panelists were asked to review the NRS EFL descriptors containing the descriptions of the achievement level categories (Sireci et al., 2008) and to match each item to a proficiency category. The NRS EFL had five different levels (see Sireci et al. (2007) and Sireci et al. (2008) for details). Two test forms with 60 items were used. Items provided to panelists were not divided into each category, but they were ordered by item difficulty parameters.

The data for the modified Angoff method included 180 items; each threshold level had 45 items. The panelists first discussed the knowledge and abilities that borderline students could have. Considering the characteristics of borderline examinees, panelists marked 1 (yes) if borderline examinees would answer an item correctly; 0 (no) otherwise. After this activity, panelists discussed their ratings in a group and resumed the procedure in the second round. Using this method, binary data for each threshold level was produced.

Analysis

Two different G-theory designs were used: a p x ($i : f$) design for the IDM method data, where p indicates a panelist facet, i denotes items, and f indicates test forms, and a ($p : g$) x i design for the modified Angoff method data, where g stands for groups of panelists. As in Lee and Lewis (2008), rounds were not considered as a facet. However, the results for two rounds were separately analyzed and compared.

To perform the G-study and D-study analyses, the computer program GENOVA (Crick & Brennan, 1983; Brennan, 2001) was used. As a result of the G-study, variance components for each facet were produced. Using the IDM method data, the G-study analysis was conducted with sample sizes of n_p=10 (math), n_p= 11 (reading), n_f = 2, and n_i = 60. With the modified Angoff method data, the sample sizes for the G-study were n_p=12, $n_{p:g}$=6, and n_i=45. In the D-study, these numbers were increased or decreased to examine possible reduction in the absolute standard error of measurement (SEM).

RESULTS

Results from the IDM Method with the p x ($i : f$) design

The percentages of variance components (VCs) of math and reading data are presented in Tables 1 and 2. Overall patterns of VCs were consistent across levels and rounds. The largest VC estimates were the residual effect $\hat{\sigma}^2(pi{:}\,f)$ for all levels and rounds, which contributed more than half of the total variance, except for one case. The second largest VC estimates were for items nested within test forms, which accounted for about 30-50% of the total variance. It is not surprising that relatively small percentages of the total variance were explained by test forms

because two forms were created to be parallel with respect to item pool representation and difficulty (Sireci et al., 2008). Variability across panelists was also small for both rounds.

TABLE 1
Percentages (%) of Variance Explained by Each Component in the Math Test

Variance Component	Level 1/Level 2		Level 2/Level 3		Level 3/Level 4		Level 4/Level 5	
	Round1	Round2	Round1	Round2	Round1	Round2	Round1	Round2
$\hat{\sigma}^2(p)$	0.6	0.8	1.1	0.8	0.0	0.3	1.2	1.2
$\hat{\sigma}^2(f)$	0	0	1.4	2.8	0	0.9	0	0
$\hat{\sigma}^2(i:f)$	42.3	49.7	33.7	38.6	30.7	34.7	34.6	37.9
$\hat{\sigma}^2(pf)$	0	0	0	0	0.4	0	0.8	0.8
$\hat{\sigma}^2(pi:f)$	57.1	49.5	63.8	57.8	68.9	64.1	63.4	60.1

Note: p = panelists; f = forms; i = items.

TABLE 2
Percentages (%) of Variance Explained by Each Component in the Reading Test

Variance Component	Level 1/Level 2		Level 2/Level 3		Level 3/Level 4		Level 4/Level 5	
	Round1	Round2	Round1	Round2	Round1	Round2	Round1	Round2
$\hat{\sigma}^2(p)$	0.2	0	0.1	0.2	1.0	1.3	2.6	1.7
$\hat{\sigma}^2(f)$	0	0	0.1	0	0	0	0	0
$\hat{\sigma}^2(i:f)$	25.9	35.3	19.8	32.1	30.8	12.7	20.8	21.8
$\hat{\sigma}^2(pf)$	0	0.8	0.2	0	0.8	0.5	0.2	0.1
$\hat{\sigma}^2(pi:f)$	73.9	63.9	79.8	67.7	67.4	85.5	76.4	76.4

Absolute standard errors of measurement using different numbers of panels, test forms, and items within the form are displayed in Figures 1 and 2. Since the variability coming from the item nested within the form facet was relatively large, lengthening the test proved more effective than having more panels or test forms. In other words, regardless of varying the number of panels or forms, doubling the number of items from 30 to 60 decreased the amount of error the most. In addition, compared to the results in the math test, the reading data results showed less absolute error. In general, however, the pattern seemed to be consistent regardless of the subjects, whether math or reading. The results for the first and second rounds showed the same pattern; thus, only the first round results are presented.

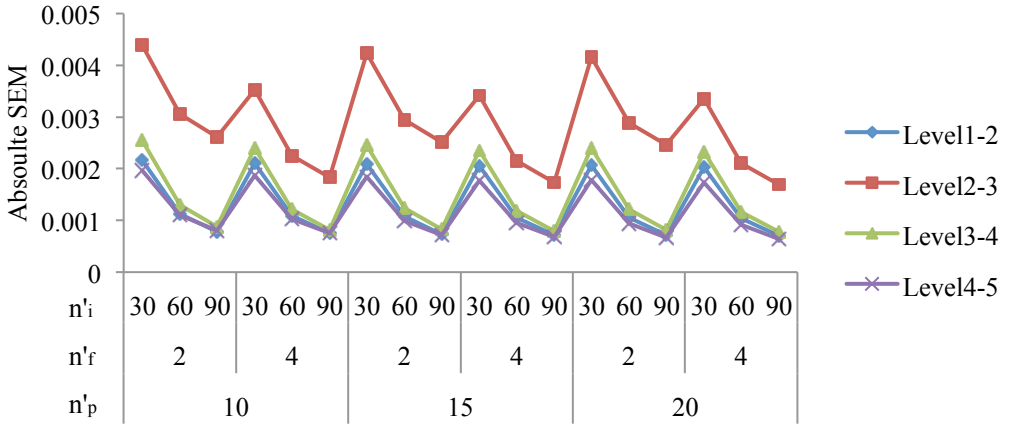

Figure 1. Absolute standard error of measurement in the Math test with the IDM method
Note: n'_i = number of items; n'_f = number of forms; n'_p = number of panelists.

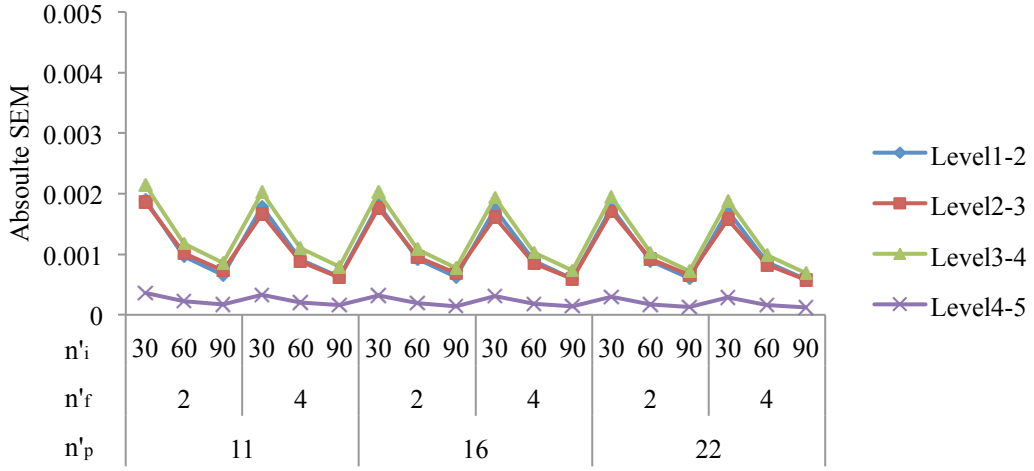

Figure 2. Absolute standard error of measurement in the Reading test with the IDM method

Results from the Modified Angoff Method with the (*p* : *g*) x *i*

The percentages of estimated variance components for the modified Angoff method using the (*p* : *g*) x *i* design are presented in Table 3. Either groups or panelists nested within the group showed the largest variance. For thresholds 1 and 4, the variance component estimates for groups were 0, but the largest VC estimates came from panelists nested within groups. In thresholds 2 and 3, the largest VC estimates were the group effect, and the second largest VC estimates were the effect of the person nested within a group. The next largest VC estimates were the item effect. Variability from group and item interaction was small across all levels.

Table 4 shows the mean probabilities indicating the average chance of examinees at the threshold correctly answering an item. The mean probability that examinees at the borderline between level 1 and level 2 answer items correctly was approximately 0.077. The same interpretation could be made for the other levels.

TABLE 3
Percentages (%) of Variance Explained by Each Component in the Reading Test with the Modified Angoff Method

Variance Component	Level 1/Level 2 (Threshold 1)	Level 2/Level 3 (Threshold 2)	Level 3/Level 4 (Threshold 3)	Level 4/Level 5 (Threshold 4)
$\hat{\sigma}^2(g)$	0	56.8	65.4	0
$\hat{\sigma}^2(p{:}\,g)$	49.1	27.0	18.4	79.6
$\hat{\sigma}^2(i)$	45.2	12.8	15.6	18.3
$\hat{\sigma}^2(gi)$	0	2.6	0.1	0.7
$\hat{\sigma}^2(p{:}\,gi)$	5.7	0.8	0.5	1.4

Note: g = groups of panelists; p = panelists; i = items.

TABLE 4
Mean Probability obtained from the Modified Angoff Method

Level	Level 1/Level 2 (Threshold 1)	Level 2/Level 3 (Threshold 2)	Level 3/Level 4 (Threshold 3)	Level 4/Level 5 (Threshold 4)
Probability	0.077	0.222	0.531	0.622

Figure 3 shows the absolute SEM. The results show the increase or decrease in the amount of error due to changing the number of each effect. The number of panels and groups was either doubled or halved from the original number. For levels 1-2 and 4-5, since the VC estimates for the group effect was zero, the values of having either 3 panels and 4 groups or 6 panels and 2 groups were identical. For level 2-3 and level 3-4, having 4 groups with 3 panels in each group showed a smaller absolute SEM than having 2 groups with 6 panels. The effect of having more groups reduced the absolute SEM more effectively than having more items because $\hat{\sigma}^2(i)$ was smaller than $\hat{\sigma}^2(p{:}\,g)$ and $\hat{\sigma}^2(g)$ in thresholds 2 and 3. Moreover, increasing the total number of panelists did not reduce errors compared to increasing the number of groups. As expected, compared to the middle levels including thresholds 2 and 3, thresholds 1 and 4 presented less error throughout all conditions.

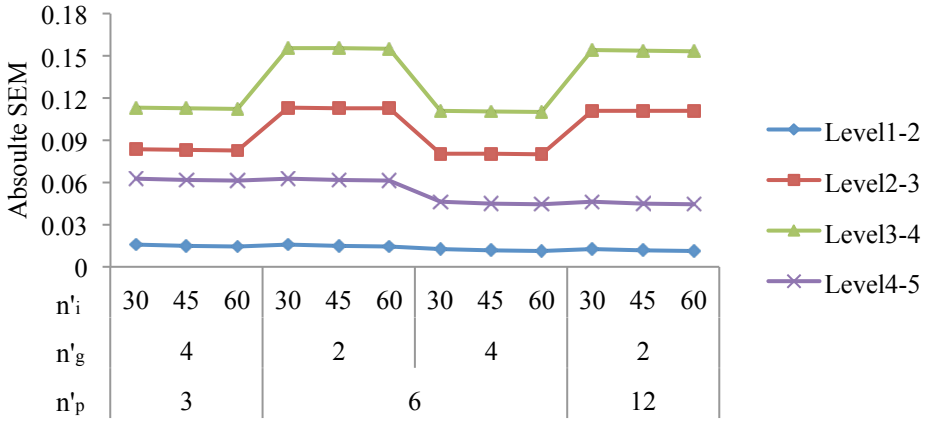

Figure 3. Absolute standard error of measurement in the Reading test with the modified Angoff method
Note: n'_i = number of items; n'_g = number of groups; n'_p = number of panelists.

DISCUSSION

This study applied the G-theory framework to MAPT standard setting studies to investigate major sources of error and find cost- and time- effective sample sizes for panels, items, forms, and groups. Two different standard setting methods were implemented in MAPT reading and math. First, the IDM method with the p x ($i : f$) design was applied. Instead of assigning people to small groups, panelists participated in a whole group discussion which resulted in small variability across panels. However, the interaction effect of panels and items nested within forms constituted a large portion of the total variance because of a relatively large magnitude of variability across items. Regardless of varying the number of panels or forms, using more items proved more effective in reducing absolute error variance than recruiting more panels or using more forms. Thus, considering cost and time, using one pair of parallel forms and having around 10 panelists seems to be reasonable. Adding more items is more efficient in reducing the absolute SEM, which can be considered in future standard setting studies. Second, the modified Angoff method with the ($p : g$) x i design was applied. When the number of panelists per group (n'_p=6, n'_g=2) or the number of groups (n'_p=3, n'_g=4) was doubled, the total number of panelists remained the same (12). However, the absolute SEM diminished when the number of groups (rather than the number of panelists per group) was doubled. Thus, doubling the number of groups can be considered more efficient in reducing the absolute SEM.

Several limitations in this study should be noted. With the results provided in this study, it was not possible to make direct inferences on the standard error of cutscores because of the scale difference. In addition, no criterion was available to compare VC estimates or absolute SEM. Only relative comparisons within the same method were made. The results would vary depending on the composition of factors such as panelists, test items, etc. The implications for this study cannot be generalized to all situations. Through this type of analysis, however, researchers are able to understand the sources of variability in the context of standard setting. When the standards need to be re-evaluated after a certain period of time, this type of analysis with the previous data would help determine the most efficient sample size by minimizing error within the budget. Due to limitations in time, human resources, and budget, sample sizes for

certain facets cannot be increased indefinitely. For instance, the number of panelists, groups, forms, and items cannot be increased unconditionally in order to reduce absolute SEM. In conclusion, sample size should be carefully chosen considering the cost, while maximizing efficiency.

REFERENCES

Brennan, R. L. (1995). Standard setting from the perspective of generalizability theory. In *Proceedings of the joint conference on standard setting for large-scale assessments* (Volume II). Washington, DC: National Center for Education Statistics and National Assessment Governing Board.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brennan, R. L., & Lockwood, R. E. (1980).A comparison of the Nedelsky and Angoff cutting score procedures using Generalizability theory. *Applied Psychological Measurement*, *4*, 219–240.

Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education*, *12*, 151–65.

Chang, L., & Hocevar, D. (2000). Models of generalizability theory in analyzing existing faculty evaluation data. *Applied Measurement in Education, 13*, 255–75.

Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2009). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement in Education, 22*, 1–21.

Clauser, J. C., Margolis, M. J., & Clauser, B. E. (2012, April). *An examination of the replicability of Angoff standard setting results within a Generalizability theory framework*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, BC.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system [Computer software and manual]*. University of Iowa.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53–88). Mahwah, NJ: Erlbaum.

Kane, M. T., & Wilson, J. (1984). Errors of measurement and standard setting in mastery testing. *Applied Psychological Measurement, 8*, 107–115.

Lee, G., & Lewis, D. M. (2008). A generalizability theory approach to standard error estimates for bookmark standard setting. *Educational and Psychological Measurement, 68,* 603-620.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.). *Standard setting: Concepts, methods, and perspectives* (pp.119-157). Mahwah. NJ: Erlbaum.

Sireci, S. G., Baldwin, P., Martone, D., & Han, K. T. (2007). *Establishing Achievement Levels on a Multi-Stage Computerized-Adaptive Test: An Application of the Item Descriptor Matching Method*. Center for Educational Assessment Research Report No. 618. Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

Sireci, S. G., Baldwin, P., Zenisky, A. L., Kaira, L., Shea, C. L., Han, K. T., Deng, N., Delton, J., & Hambleton, R. K. (2008). *Massachusetts Adult Proficiency Tests technical manual* Version 2. Center for Educational Assessment Research Report No. 677. Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

Yin, P., & Sconing, J. (2008). Estimating standard errors of cut scores for item rating and mapmark procedures. *Educational and Psychological Measurement, 68*, 25-41.

KAERA Research Forum, *1(1)*, 55-71

RESEARCH ARTICLE

Dealing with Measurement Error in Estimating a Cross-level Interaction: Nonlinear Multilevel Latent Variable Modeling Approach with a Metropolis-Hastings Robbins-Monro Algorithm

Ji Seung Yang
University of Maryland

Li Cai
University of California, Los Angeles

The main purpose of this study is to improve estimation efficiency in obtaining full-information maximum likelihood (FIML) estimates of cross-level interactions in the framework of a nonlinear multilevel latent variable model by adopting the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b). Results indicate that the MH-RM algorithm can produce FIML estimates and their standard errors efficiently for a cross-level interaction model that requires high dimensional integration. Simulations, with various sampling and measurement structure conditions, were conducted to obtain information about the performance of nonlinear multilevel latent variable modeling compared to traditional hierarchical linear modeling. Results suggest that nonlinear multilevel latent variable modeling can more properly estimate and detect a cross-level interaction effect than the traditional approach. As an empirical illustration, a subset of data extracted from The Programme for International Student Assessment (PISA, 2000; OECD, 2000) was analyzed.

Keywords: Cross-level interaction, Multilevel latent variable modeling, MH-RM Algorithm.

In educational research, outcomes and predictors of research interest are often measured by sets of items. To come up with a single score that represents a level of a certain variable or psychological construct (e.g., academic achievement, self-concept, and locus of self-control) the item-level raw scores (e.g., Likert scales) are often summed or averaged and the single score is used in statistical modeling. However, measurement error poses significant challenges to the application of standard multilevel models to impact evaluation studies. Since the summed or averaged item scores are not free from measurement error, using error-contaminated predictors in statistical modeling causes attenuated correlation coefficients and regression coefficients (Spearman, 1904).

When the data collection is extended to a multilevel context, which requires a hierarchical linear model (HLM) analysis, the measurement error issue becomes dually burdensome because the error-contaminated predictor values are often aggregated to the upper level to form the upper-

level predictors. Accordingly, there are two methodological issues in estimating regression coefficients in multilevel models when predictors are particularly measured by sets of categorical variables: The first one is related to the attenuated coefficient estimates due to measurement error in predictors (Spearman, 1904), and the other is biased parameter estimates due to sampling error associated with aggregating level-1 variables to form level-2 variables by simply averaging the values (Raudenbush & Bryk, 2002, chap.3). Therefore, two regression coefficients at level-1 and level-2 tend to be attenuated when summed or averaged scores are used as predictors.

To handle measurement error and sampling error more properly, multilevel latent variable modeling has been suggested as an alternative to traditional methods (e.g., Lüdtke et al., 2008; Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Marsh et al., 2009). For example, Lüdtke et al. (2008) proposed a multilevel latent variable modeling framework for contextual analysis. Lüdtke et al. (2008) examined the relative bias in contextual effect estimates when the traditional HLM is used under different data conditions. The results showed that the relative percentage bias of contextual effect was less than 10% across varying data conditions when a multilevel latent variable model was used. On the other hand, the relative percentage bias of contextual effect was up to 80% when the traditional HLM model was used. However, the traditional HLM can yield less than 10% relative bias under favorable data conditions—that is, when level-1 and level-2 units exceed 30 and 500, respectively, and when there is substantial intra-class correlation (ICC) in the predictor (e.g., 0.3). While the manifest variables are limited to only continuous variables in Lüdtke et al. (2008), multiple categorical variables are used as manifest variables for both latent predictor and outcome variables in the current study.

While nonlinear multilevel latent variable modeling can deal with measurement and sampling error properly, this approach presents significant computational difficulties with categorical manifest variables. Standard approaches such as numerical integration (e.g., adaptive quadrature) or Markov chain Monte Carlo (MCMC; e.g., Gibbs Sampling) based estimation methods have important limitations that make them less practical for routine use, because their computational efficiency drops dramatically when the dimensionality is high. Lüdtke et al. (2011) also reported the occurrence of unstable estimates. The model has difficulty in converging when sample size is small and the intraclass correlation coefficient (ICC) in a predictor is small. Therefore, further research efforts are needed to improve estimation of contextual effect in the nonlinear multilevel latent variable modeling framework.

The main objective of this study was to develop a more efficient estimation method for cross-level interactions in the nonlinear multilevel latent variable modeling framework, by adopting the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b). Computational efficiency and parameter recovery were assessed in a comparison with an existing EM algorithm using adaptive Gauss-Hermite quadrature for numerical integration (e.g., Mplus; Muthén & Muthén, 2008). Another objective was to find, through a simulation study, how much measurement error and sampling error can influence the cross-level interaction estimates under different conditions. The results provide the rationale for using computationally demanding nonlinear multilevel latent variable models. The last objective of the proposed study was to provide an empirical illustration by applying nonlinear multilevel latent variable models to real data that contain more complex measurement structures and unbalanced data. A subset from The Programme for International Student Assessment (PISA; Adams & Wu, 2002) was analyzed to illustrate a cross-level interaction model.

The particular contextual effect of interest in this study is one that occurs when a group-level characteristic of interest is measured by individual-level characteristics, and the individual-level characteristics are measured by categorical manifest variables. The parameter of interest in

this study is a cross-level interaction that captures the influence of contextual variables on within-group slopes.

NONLINEAR MULTILEVEL LATENT VARIABLE MODEL

Structural Models

The traditional HLM defines a cross-level interaction γ_{11} as follows:

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}(X_{ij} - X_{\cdot j}) + r_{ij}, \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}(X_{\cdot j} - X_{\cdot\cdot}) + u_{0j}, \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(X_{\cdot j} - X_{\cdot\cdot}) + u_{1j}. \end{aligned} \tag{1}$$

In Equation (1), Y_{ij} and X_{ij} denote outcome and predictor values of student i in school j, respectively. Y_{ij} and X_{ij} are typically constructed by summing item scores on self-report responses. The random effects r_{ij}, u_{0j}, and u_{1j} are assumed to be normally distributed with zero means and variances (σ^2 and $\boldsymbol{\tau}$). γ_{11} is the parameter of research interest, which is the regression coefficient for the cross-level interaction term between level-1 and level-2 predictors when the equations are collapsed to a single level by substituting β_{0j} and β_{1j} as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(X_{\cdot j} - X_{\cdot\cdot}) + \gamma_{10}(X_{ij} - X_{\cdot j}) + \gamma_{11}(X_{\cdot j} - X_{\cdot\cdot})(X_{ij} - X_{\cdot j}) + r_{ij} + u_{0j} + u_{1j}(X_{ij} - X_{\cdot j}) \tag{2}$$

In a nonlinear multilevel latent variable model, instead of using Y_{ij} and X_{ij} that are observed variables, we substitute them with latent variables η_{ij} and ξ_{ij} for individual i in group j. Those latent variables are connected to manifest variables through measurement models. For notational simplicity, latent individual deviations from latent group means ($\xi_{ij} - \xi_{\cdot j}$) can be defined as δ_{ij}, and group mean deviations from the latent grand mean ($\xi_{\cdot j} - \xi_{\cdot\cdot}$) can be defined as $\delta_{\cdot j}$. Then Equation (1) translates into the following cross-level interaction effect model:

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \beta_{1j}\delta_{ij} + r_{ij}, \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\delta_{\cdot j} + u_{0j}, \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}\delta_{\cdot j} + u_{1j}. \end{aligned} \tag{3}$$

In Equation (3), γ_{11} is the parameter of research interest, which is the regression coefficient for the cross-level interaction term between level-1 and level-2 predictors, taking into measurement error and sampling error into account.

Measurement Models

The measurement models define the relationship between observed (manifest) variables and latent variables. For simplicity, only the measurement models of level-1 latent predictor variable ξ_{ij} will be described in this section, since the measurement models for other variables such as the latent outcome η_{ij} follow the same principles.

For example, when manifest variables are graded response variables with multiple categories, Samejima's (1969) model can be utilized. Let $x_{ijl} \in \{0, 1, 2, ..., K_l - 1\}$ be an element of ith individual's response in jth group to lth item that has K_l ordered categories. Then the logistic conditional cumulative response probability for each category is listed as follows:

$$\begin{aligned}
P_\theta(x_{ijl} \geq 0|\xi_{ij}) &= 1, \\
P_\theta(x_{ijl} \geq 1|\xi_{ij}) &= \frac{1}{1+\exp[-(b_{1,l} + a_l\xi_{ij})]}, \\
P_\theta(x_{ijl} \geq 2|\xi_{ij}) &= \frac{1}{1+\exp[-(b_{2,l} + a_l\xi_{ij})]}, \\
&\vdots \\
P_\theta(x_{ijl} \geq K_l - 1|\xi_{ij}) &= \frac{1}{1+\exp[-(b_{K_l-1,l} + a_l\xi_{ij})]}.
\end{aligned} \tag{4}$$

The category response probability is defined as the difference between two adjacent cumulative probabilities:

$$P_\theta(x_{ijl} = k|\xi_{ij}) = P_\theta(x_{ijl} \geq k|\xi_{ij}) - P_\theta(x_{ijl} \geq k+1|\xi_{ij}), \tag{5}$$

where $P_\theta(x_{ijl} \geq K_l|\xi_{ij})$ is zero. χ_k is an indicator function in which χ_k is 1 if $x_{ijl} = k$, or 0 otherwise. The conditional density for x_{ijl} follows a multinomial with trial size 1 in K_l categories:

$$f_\theta(x_{ijl}|\xi_{ij}) = \prod_{k=0}^{K_l-1} P_\theta(x_{ijl} = k|\xi_{ij})^{\chi_k(x_{ijl})}. \tag{6}$$

Any item response model can be utilized for the measurement model depending on situations, and the observed and complete data likelihoods for the model are needed to estimate the model parameters.

METROPOLIS-HASTINGS ROBBINS-MONRO ALGORITH FOR NONLINEAR MULTILEVEL VARIABLE MODEL

An MH-RM algorithm was initially proposed by Cai (2008) for nonlinear latent structure analysis with a comprehensive measurement model, and the application of the algorithm has been expanded to further measurement and statistical models (e.g., Cai, 2010a, 2010b). The MH-RM algorithm was motivated by Fisher's Identity (Fisher, 1925), which proved that the gradient of the observed likelihood is the expectation of the gradient of the complete likelihood. While maximizing the observed likelihood, denoted as $L(\theta|\mathbf{Y}_o)$, involves high-dimensional integrals, the complete data likelihood, denoted as $L(\theta|\mathbf{Y})$, involves a series of products of likelihoods that are fairly simple to maximize. Therefore, having plausible values of random effects and latent variables makes the estimation problem simpler. This also allows straightforward optimization of the complete data likelihood with respect to θ. However, proper imputation requires the distribution of the missing data to be conditional on the observed data. As the model is nonlinear, analytical derivation of the distribution of missing data conditional on the observed data is difficult. Nevertheless, a property of the posterior of the missing data enables us to have appropriate imputation. That is, the posterior of missing data, given observed data and a provisional θ, is proportional to the complete data likelihood. To utilize this property, Metropolis-Hastings sampler (MH; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) is adopted to produce the imputations from a Markov chain with the missing data posterior as the target. Then, the random imputations are combined into Stochastic Approximation using the Robbins-Monro algorithm (RM; Robbins & Monro, 1951).

The $(k + 1)$th iteration of the MH-RM algorithm consists of 3 steps: Stochastic Imputation, Stochastic Approximation, and Robbins-Monro Update.

Step 1. Stochastic Imputation

Draw m_k sets of missing data, which are the random effects and latent variables, from a Markov chain that has the distribution of missing data conditional on observed data as the target. Then, m_k sets of complete data are as follows:

$$\left\{\mathbf{Y}_j^{k+1}; j = 1, ..., m_k\right\} \tag{7}$$

Step 2. Stochastic Approximation

Using Fisher's Identity, a Monte Carlo approximation to $\nabla_\theta l(\boldsymbol{\theta}^k|\mathbf{Y}_o)$ can be computed as the sample average of complete data gradients. We also compute a recursive approximation of the conditional expectation of the information matrix of the complete data log-likelihood. For simplicity, let $\mathbf{s}(\boldsymbol{\theta}|\mathbf{Y})$ stand for $\nabla_\theta l(\boldsymbol{\theta}|\mathbf{Y})$, and the sample average of complete data gradients can be written as:

$$\tilde{\mathbf{s}}_{k+1} = \frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{s}(\boldsymbol{\theta}^k | \mathbf{Y}_j^{k+1}), \tag{8}$$

and $\boldsymbol{\Gamma}_{k+1}$ is

$$\boldsymbol{\Gamma}_{k+1} = \boldsymbol{\Gamma}_k + \gamma_k \left[\frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{H}(\boldsymbol{\theta}^k | \mathbf{Y}_j^{k+1}) - \boldsymbol{\Gamma}_k \right], \tag{9}$$

where $\mathbf{H}(\boldsymbol{\theta}|\mathbf{Y})$ is the complete data information matrix, which is -1 times the second derivative matrix of the complete data log-likelihood.

Step 3. Robbins-Monro Update

Now new parameters are estimated through the following update:

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \gamma_k (\boldsymbol{\Gamma}_{k+1}^{-1} \tilde{\mathbf{s}}_{k+1}). \tag{10}$$

The whole iteration process is composed of three stages: an initial stage in which parameters are not updated (M1), a constant gain stage in which parameters are updated with a constant gain (M2), and the decreasing gain stage in which parameters are updated with a decreasing constant gain so that they stop oscillating around the MLE (M3). The iterations can be stopped upon convergence when the changes in parameter estimates are sufficiently small. Cai (2008) verified the asymptotic behaviors of MH-RM in time and that it converges to MLE. For further details about the algorithm itself, readers can refer to Cai (2008, 2010a, 2010b).

Approximation to the Observed Information Matrix

One of the benefits of using the MH-RM algorithm is that the observed data information matrix can be recursively approximated as a byproduct of the iterations. The inverse of the observed data information matrix becomes the large-sample covariance matrix of parameter estimates. The square root of the diagonal elements are the standard errors. Another practical option for approximating the observed information matrix is a direct application of Louis's (1982) approach, in which the score vector and the conditional expectation are approximated directly after they converge. In this study, the first method is called *recursively approximated standard errors* and the latter is called *post-convergence approximated standard errors*. Based on preliminary study results, the reported standard errors in this paper are post-convergent approximated standard errors which are slightly underestimated than the recursively approximated standard errors but more easily to satisfy the second order matrix convergence criteria.

SIMULATION STUDIES

Simulation Study 1: Comparison of Estimation Algorithms

The first simulation study was to examine the parameter recovery and standard errors when an MH-RM algorithm is implemented in comparison to those from an existing EM algorithm.

Method. The data-generating and fitted models followed Equation (3) for a cross-level interaction model. The simulated data are balanced in that the number of level-2 units (*ng*) is 100 and the number of level-1 units per group (*np*) is 20. The generating ICC value for the latent predictor was 0.3.

For the measurement model, five dichotomously scored manifest variables were generated for each latent trait (i.e., η, and ξ) using a 2-PL model. The item parameters were the same across levels, representing cross-level measurement invariance.

One hundred data sets were generated with the same parameters but with 100 different random seeds for each model. The first 10 data sets were analyzed using two methods: an MH-RM algorithm implemented in R (R Core Team, 2012) and an adaptive quadrature EM approach implemented in Mplus (Muthén & Muthén, 2010). Then the other 90 data sets are all analyzed using the MH-RM algorithm.

The MH-RM algorithm convergence criterion was 5.0×10^{-5} and the maximum numbers of iterations for each stage were $M1 = 100$, $M2 = 800$, and $M3 = 800$. To calculated post-convergence approximated standard errors, 100 to 800 samples were used. The convergence rates at the given number of iterations were 52%.

Results. The generating values and the corresponding estimates from analyzing the first simulated data set using different algorithms are summarized in Table 1. The number of quadrature points for the EM algorithm makes some noticeable differences in the mean point estimates as well as the standard errors.

Efficiency of the MH-RM algorithm compared to the EM algorithm was more prominent for this cross-level interaction model, even as it is still in R. Using Mplus, even with 8 processors, the estimation took more than 1 hour and 30 minutes, while it took similar or even shorter time for the MH-RM algorithm implemented in R. When 1 processor was used, it took about 4 to 5 hours to yield a result using Mplus. This difference is remarkable considering that R does not have support for multi-processors.

For further analysis, more simulated data sets were analyzed by applying the MH-RM algorithm, and the generating values and corresponding estimates are summarized in Table 2. The largest relative bias of the parameter estimates for both measurement and structural parts is less than 10%. Means of standard error estimates and Monte Carlo standard deviations of point estimates are reasonably compatible; however, underestimation of standard errors for threshold estimates was consistent, indicating that the post-convergence approximation approach can be chosen for efficiency reasons, but with a cost in accuracy. It is notable that the standard error of the cross-level interaction is quite large ($E(\hat{\gamma}_{11})$ =0.46, E(SE)=0.27) particularly compared to those of other parameters, indicating the variability of cross-level interaction across samples is large even under the fairly desirable sampling condition. Given the iteration conditions, only 26 of 50 replications converged within the specified number of iterations. For this condition, the cause of low convergence rate was mostly due to the approximation of observed data information matrix rather than point estimates themselves. Either allowing larger numbers of iterations or achieving more efficient approximation of the observed data information matrix would help the

convergence rate increase. As a trial, 1000 iterations were tried, and this could increase the convergence rate up to 78% for this condition.

TABLE 1
Generating values and estimates for a cross-level interaction model
(N=2,000, ng=100, np=20, 1st simulated data set)

		EM (5qp)		EM (8qp)		MH-RM	
	Θ	$E(\hat{\theta})$	$E\{se(\hat{\theta})\}$	$E(\hat{\theta})$	$E\{se(\hat{\theta})\}$	$E\hat{\theta})$	$E\{se(\hat{\theta})\}$
Structural parameters							
γ_{01}	1.00	1.86	0.25	1.35	0.22	1.44	0.22
γ_{10}	0.50	1.94	0.15	0.63	0.13	0.63	0.05
γ_{11}	0.50	1.27	0.45	0.83	0.29	0.83	0.06
τ_{00}	1.00	0.85	0.11	0.88	0.12	0.90	0.18
τ_{11}	1.00	0.78	0.33	0.83	0.25	0.79	0.16
τ_{01}	0.50	0.96	0.15	0.49	0.12	0.49	0.11
var(ξ.j)	0.43	0.40	0.02	0.39	0.05	0.39	0.07
Measurement parameters							
a_{x1}	0.80	0.78	–	0.78	–	0.78	0.08
a_{x2}	1.00	1.40	0.14	0.96	0.14	0.96	0.07
a_{x3}	1.20	2.05	0.19	1.41	0.19	1.41	0.12
a_{x4}	1.40	2.37	0.21	1.62	0.21	1.63	0.18
a_{x5}	1.60	2.51	0.24	1.69	0.25	1.71	0.12
a_{y1}	0.80	0.79	0.00	0.79	0.00	0.79	0.05
a_{y2}	1.00	0.95	0.11	0.93	0.11	0.93	0.06
a_{y3}	1.20	1.17	0.11	1.15	0.12	1.16	0.07
a_{y4}	1.40	1.00	0.14	0.98	0.15	1.22	0.08
a_{y5}	1.60	1.43	0.18	1.40	0.19	1.51	0.09
c_{x1}	-0.80	-0.68	0.06	-0.73	0.07	-0.74	0.05
c_{x2}	0.00	0.10	0.08	0.10	0.08	0.09	0.05
c_{x3}	1.20	1.43	0.11	1.43	0.12	1.41	0.09
c_{x4}	-0.70	-0.52	0.11	-0.51	0.12	-0.53	0.08
c_{x5}	0.80	1.11	0.13	1.10	0.14	1.09	0.08
c_{y1}	-0.80	-0.72	0.09	-0.73	0.11	-0.73	0.06
c_{y2}	0.00	0.03	0.11	0.04	0.13	0.03	0.06
c_{y3}	1.20	1.26	0.14	1.26	0.16	1.26	0.08
c_{y4}	-0.70	-0.53	0.14	-0.52	0.16	-0.52	0.07
c_{y5}	0.80	0.96	0.17	0.96	0.20	0.96	0.08
Efficiency							
8 processors		15 min		100 min		60min	
1 processor		40 min		4hour 40 min			

Note: θ = Generating values; $E(\hat{\theta})$ = mean of point estimates; $E\{se(\hat{\theta})\}$ = mean of estimated SEs (post-convergence approximated SEs); a = item slope parameter; c = item threshold parameter; qp = number of quadrature points used in estimation. Mplus does not allow standardized factor identification option; therefore, anchoring the first factor loading option was used to estimate the model and the results are transformed to make the estimate comparable. The differences are particularly prominent in the structural parameters and the slopes of predictor-side indicators, as

within-level variance estimates of the predictor were different across the number of quadrature points being used. However, the results from MH-RM algorithm are closer to the 8-quadrature-points results, indicating that reducing the number of quadrature points for a higher dimensional model is not desirable.

TABLE 2

Generating values and estimates for a cross-level interaction model using MH-RM algorithm (N=2,000, ng=100, np=20, 26/50 converged)

	Θ	$E(\hat{\theta})$	$E\{se(\hat{\theta})\}$	$SD(\hat{\theta})$
Structural parameters				
γ_{01}	1.00	1.07	0.18	0.21
γ_{10}	0.50	0.55	0.07	0.14
γ_{11}	0.50	0.46	0.27	0.19
τ_{00}	1.00	1.06	0.29	0.17
τ_{11}	1.00	1.05	0.28	0.27
τ_{01}	0.50	0.50	0.15	0.12
$var(\xi.j)$	0.43	0.43	0.07	0.09
Measurement parameters				
a_{x1}	0.80	0.78	0.08	0.06
a_{x2}	1.00	0.98	0.08	0.08
a_{x3}	1.20	1.23	0.11	0.09
a_{x4}	1.40	1.37	0.12	0.14
a_{x5}	1.60	1.59	0.18	0.12
a_{y1}	0.80	0.77	0.06	0.06
a_{y2}	1.00	0.97	0.07	0.06
a_{y3}	1.20	1.19	0.11	0.06
a_{y4}	1.40	1.37	0.12	0.14
a_{y5}	1.60	1.56	0.17	0.13
c_{x1}	-0.80	-0.77	0.06	0.09
c_{x2}	0.00	0.00	0.05	0.09
c_{x3}	1.20	1.21	0.08	0.12
c_{x4}	-0.70	-0.66	0.07	0.14
c_{x5}	0.80	0.78	0.08	0.14
c_{y1}	-0.80	-0.79	0.06	0.12
c_{y2}	0.00	0.00	0.06	0.15
c_{y3}	1.20	1.21	0.09	0.19
c_{y4}	-0.70	-0.67	0.08	0.23
c_{y5}	0.80	0.84	0.09	0.24
Efficiency				
60~90min				

Note. θ = Generating values; $E(\hat{\theta})$ = mean of point estimates; $E\{se((\hat{\theta})\}$ = mean of estimated SEs (post-convergence approximated SEs); a = item slope parameter; c = item threshold parameter.

Simulation Study 2: Comparison of Models

Method. The second simulation study was conducted to examine how measurement error and sampling error may influence compositional effect and cross-level interaction estimates across different conditions with both a traditional HLM model and a latent variable model.

*Simulation conditions***.** Data generation conditions varied with respect to cross-level interaction sizes (0, 0.5 or 1), sampling conditions (*ng*=100, *np*=20; *ng*=100, *np*=5; *ng*=20, *np*=20), ICC sizes (0.1 or 0.3), and measurement conditions (see Table 3). Specified conditions for sampling and ICC sizes were determined based on previous research Marsh et al. (2009), and measurement conditions reflected a short form of typical affective domain items. 50 replications were attempted.

*Analysis***.** Each data set has three sets of parameter estimates: 1) estimates from analyzing the generating values of η_{ij} and ξ_{ij} with a traditional multilevel model, which is treated as the gold standard (denoted as *G*), 2) estimates obtained by applying latent variable model (denoted as *L*), and 3) the estimates from analyzing the observed summed scores with the manifest variable approach (denoted as *M*). All of the traditional HLM analyses were conducted using an R package *nlme* (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2012).

*Statistics***.** To compare these three sets of estimates, three statistics are calculated: 1) the percentage bias of the estimate relative to the magnitude of generating value, 2) the observed coverage of the 95% confident interval (CI) for true value, and 3) the observed power to detect the effect of interest as significant.

It should be noted that the regression coefficient estimates from the observed summed score analysis using a traditional multilevel model are not on the same scales as those obtained using the latent variable approach. To make the coefficient estimates more comparable, the estimates from traditional model approach were standardized by multiplying the parameter estimates by the ratio of standard deviation of the predictor to the standard deviation of the outcome.

TABLE 3
Conditions of measurement models and generating values for item parameters

Measurement Model 1		
Condition	Slope	Intercept
	ξ_{ij} indicators X1~X5 (2PL)	η_{ij} indicators Y1~Y5 (2PL)
X1, Y1	0.8	-1.0
X2, Y2	1.0	0
X3, Y3	1.2	1.0
X4, Y4	1.4	-0.5
X5, Y5	1.6	0.5
Measurement Model 2		
	ξ_{ij} indicators X1~X5 (GR, K=5)	η_{ij} indicators Y1~Y5 (GR, K=5)
X1, Y1	0.8	-1, 0, 1, 2
X2, Y2	1.0	-1, 0, 1, 2
X3, Y3	1.2	-1, 0, 1, 2
X4, Y4	1.4	-1, 0, 1, 2
X5, Y5	1.6	-1, 0, 1, 2

Results. The relative percentage bias in $\hat{\gamma}_{11}$ across simulated data conditions is summarized in Figure 1. First, when generating values are analyzed, bias can be as small as about 2% when the sampling condition is favorable and ICC is large enough. However, the bias can be as large as about 40% even when generating values are analyzed when the ICC is small and the number of groups sampled is 25. While the traditional approach yields more than 75% underestimation across conditions and reached almost 100% when a small ICC is combined with limited sample conditions, the bias in $\hat{\gamma}_{11}$ from the latent variable model analysis was smaller than that from the manifest variable model analysis.

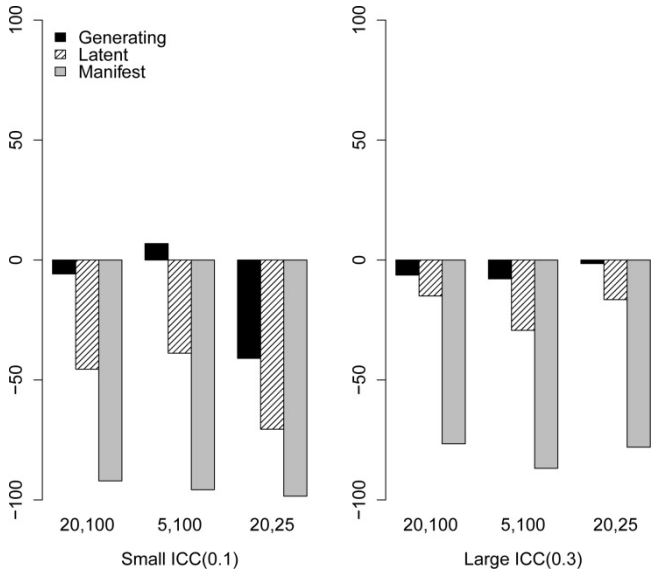

Figure 1. Relative Percentage Bias in $\hat{\gamma}_{11}$, Small CLI, MM 1.

Coverage rates for true cross-level interaction effects using 95% confidence intervals are reported in Figure 2. When generating values were analyzed, 95% confidence intervals covered the true cross-level interaction 81 to 100% of the time. When the latent variable model was applied, the coverage rates ranged from 12 to 87% depending on sampling conditions. When the number of sampled groups was small, the confidence intervals hardly captured the true values, even with the latent variable modeling approach. However, these coverage rates were still much higher than those from the traditional model approach. As bias in estimates was big and the standard error estimates were small in the traditional model approach, it was extremely rare to observe that confidence intervals actually covered the true value. Most of the coverage rates were 0.

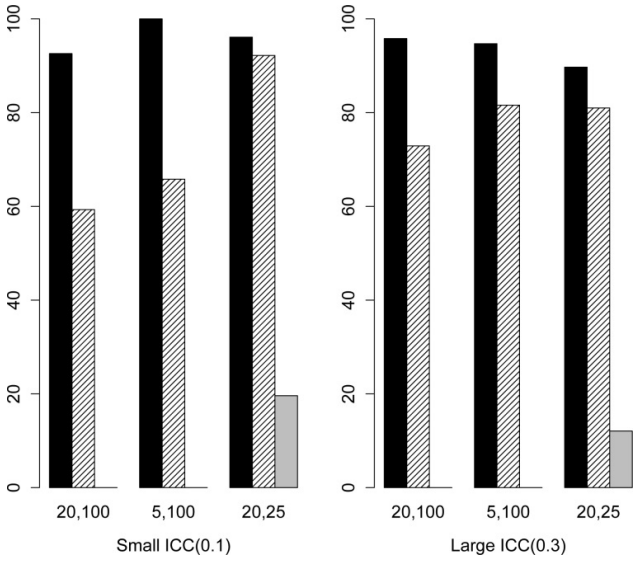

Figure 2. 95% coverage rates of $\hat{\gamma}_{11}$, Small CLI, MM 1.

Figure 3 shows the observed percentage of significant cross-level interaction across different sampling conditions and analysis models. Results from the generating value analyses are encouraging in that power can be about .80 for both large and small cross-level interactions, as long as ICC is large enough and a sufficient number of groups is sampled. However, when a small number of groups is sampled, the power can be as low as .32 for a large cross-level interaction and .06 for a small cross-level interaction. The latent variable model approach can detect cross-level interaction better than the traditional modeling approach in that the percentages of significant cross-level interactions are higher in general than those from the traditional model analysis. However, when the cross-level interaction is large and the sampling condition is favorable with large ICC, the traditional model can detect the effect slightly more frequently than the latent variable modeling approach. However, it should be noted that the CI's do not cover the true value in this case, even though the traditional model can detect the existence of the cross-level interaction. It is notable that the power of the traditional model decreases dramatically when either ICC or the number of people per group is small.

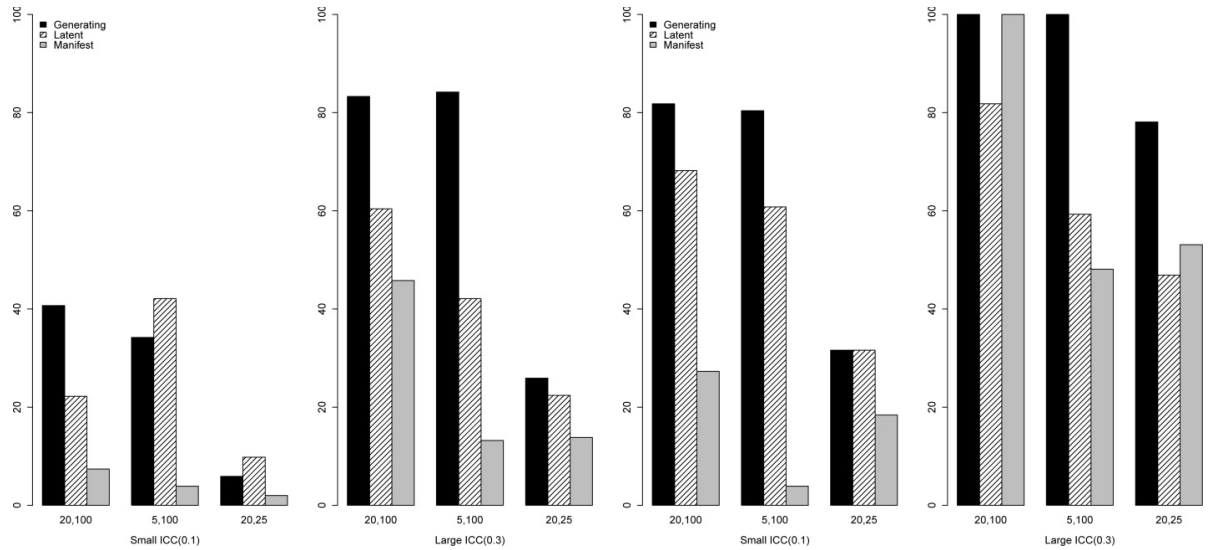

Figure 3. Percentage of significant cross-level interaction effect, Small (first two plots) and Large (last two plots) CLI, MM 1.

EMPIRICAL APPLICATION: CO-OPERATIVE LEARNING PREFERENCE AND READING LITERACY

Data

For this cross-level interaction model analysis, a subset of PISA 2000 was extracted and analyzed. The data were collected in Korea, and students who were administered booklets 8 and 9 for reading literacy were used in this analysis. In the process of data cleaning, 4 reading items were dropped, since all item responses were zero. 29 item responses (3 polytomous items with 0, 1, and 2 categories and 26 dichotomously scored items) of 1,103 students in 143 schools were analyzed. These 29 items are the indicators for the latent predictor variable. The number of students within a school ranged from 1 to 8, which can be considered a small number of students per group. The outcome variable, *co-operative learning preference*, was measured by four items (CC02Q02, CC02Q08, CC02Q19, and CC02Q22). Each item has a Likert-type scale, ranging from 1 (disagree) to 4 (agree).

Results

The structural parameter estimates from the multilevel latent variable model analysis (EM algorithm and the MH-RM algorithm) and traditional multilevel model analysis are reported in Table 4. In general, positive within- and between-level coefficients ($\hat{\gamma}_{10}$ and $\hat{\gamma}_{01}$) were found, indicating that the level of co-operative learning preference and reading literacy is positively associated. However, none of these were statistically significant when the MH-RM algorithm was applied, and only the between-level coefficient was significant at a $p < .05$ level when the EM algorithm was applied, which is also different from the traditional HLM analysis in that both coefficients are statistically different from 0 due to the small standard errors.

The parameter estimate of interest that captures a cross-level interaction effect was $\hat{\gamma}_{11}$, which appears to be negative in this particular example across computational algorithms and models. The negative cross-level interaction can be interpreted as the relationship between co-

operative learning preference and reading literacy is weaker in schools with higher achievement levels, indicating the slope of between two variables becomes less stiff as school-level achievement increases. If the negative cross-level interaction size is large enough, the direction of the relationship between the co-operative learning preference and reading literacy could be negative at schools where school-level reading literacy is very high. However, $\hat{\gamma}_{11}$ was not statistically different from 0 across models and computational algorithms.

With respect to computation, an 8 adaptive quadrature points estimation using Mplus did not converge, and only a 5-quadrature-point solution was available with some changes in default settings that are related to the M-step. When the MH-RM algorithm was applied, it took 18 hours to estimate, and a large number of samples (3,000) were used to calculate the observed data information.

TABLE 4.
Structural parameter estimates from PISA 2000 Korea data analysis using the cross-level interaction model

	Latent variable model						Manifest variable model		
	MH-RM			EM			EM		
Parameter θ	$\hat{\theta}$	se($\hat{\theta}$)	*t*-value	$\hat{\theta}$	se($\hat{\theta}$)	*t*-value	$\hat{\theta}$	se($\hat{\theta}$)	*t*-value
γ_{10}	0.021	0.061	0.315	0.229 (0.018)	0.149	1.538	0.066	0.019	3.339
γ_{01}	0.045	0.068	0.739	0.233 (0.032)	0.009	26.972	0.041	0.016	2.618
γ_{11}	-0.088	0.062	-1.417	-0.364 (-0.050)	0.296	-1.232	-0.004	0.019	-1.363
τ_{00}	0.021	0.005	4.556	0.002 (0.034)	0.000	3.918	0.353	0.594 (*SD*)	192.83 (χ^2)
τ_{11}	0.073	0.015	4.709	1.744 (0.060)	0.615	2.837	0.005	0.070 (*SD*)	147.04 (χ^2)
τ_{01}	-0.029	0.006	-4.517	-0.052 (-0.030)	0.016	-3.211	-0.023	0.598 (*SD*)	172.75 (χ^2)
$var(\xi_{.j})$	0.817	0.007	118.852	0.629 (0.830)	0.088	7.123	N/A	N/A	N/A
Computation time	18 hours M1=100, M2=1000, M3=1000 3000 for SE burn-in=5			8 hours 5qp,1processor Mstep iteration=5000 M convergence=0.00001					

Note. Reported standard errors for the MH-RM algorithm are obtained using the post-convergence approximated observed data information. Numbers in () are transformed point-estimates for comparison since different identification option was used form Mplus running. M1=Number of maximum iterations at initializing stage; M2=Number of maximum iterations at the constant gain stage; M3=Number of maximum iterations at the decreasing gain stage; qp=number of adaptive quadrature points.

DISCUSSION

This study is situated in the current streams of research (e.g., Goldstein & Browne, 2004; Goldstein, Bonnet, & Rocher, 2007; Kamata, Bauer, & Miyazaki, 2008) that try to develop a comprehensive unified model that benefits from both multilevel modeling and latent variable modeling by combining multidimensional IRT and factor analytic measurement modeling with the flexibility of nonlinear structural modeling in a multilevel setting. Considering that one of the most urgent needs in developing a unified model is an efficient estimation method, the current study contributes to nonlinear multilevel latent variable modeling by investigating an alternative estimation algorithm. The principles of the MH-RM algorithm and the previous study results (Cai, 2008) suggest that the algorithm can be more efficient than the existing algorithms when a model is associated with a large number of latent variables or random effects.

The main purpose of this study was to improve estimation efficiency in obtaining full-information maximum likelihood (FIML) estimates of cross-level interactions by adopting the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b). R programs (R Core Team, 2012) implementing the MH-RM algorithm were produced to fit nonlinear multilevel latent variable models. Computation efficiency and parameter recovery were assessed by comparing results with an EM algorithm that uses adaptive Gauss-Hermite quadrature for numerical integration. Results indicate that the MH-RM algorithm can obtain FIML estimates and their standard errors efficiently. While using the EM algorithm with only 8 adaptive quadrature points required about 100 minutes to estimate a cross-level interaction model, the MH-RM algorithm required about 60 minutes to have similar results. Considering the difference between an interpreted language and a compiled language in which each algorithm is implemented, even more substantial improvement in efficiency is expected if the MH-RM algorithm is written in a compiled language in the future.

The second purpose of this study was to provide information about the performance of nonlinear multilevel latent variable modeling compared to traditional HLM through a simulation study with various sampling and measurement structure conditions. Results suggest that nonlinear multilevel latent variable modeling can more properly estimate and detect a contextual effect than the traditional approach in most conditions. Substantial bias was found in the between-level coefficient and in the cross-level interaction coefficient when the traditional model is applied, notably, when the intraclass correlation (ICC) and the number of individuals per group were both small, the bias can be more than 80%, and the CIs hardly capture the true values. This is because when the ICC is small, the between-group variance is too small to be decomposed and estimated, indicating between-group variation is small and the characteristic of interest is homogenous across groups. When this issue is combined with a small number of groups or a small number of people per group, the condition exacerbates the difficulty in estimating between-group variance and yields difficulty in convergence and biased estimates.

When traditional models are used, not only bias in point estimate but also Type I error were problematic. Type I error rates of the traditional model are substantially elevated (up to 60%) in this sampling condition, indicating that the compositional effect detected by the traditional model under desirable sampling conditions could be spurious. These unacceptable Type I error rates are caused by the small standard error of between-level regression coefficient in the traditional HLM. The standard error of the between-level coefficients in HLM is influenced by the variance of between-level coefficient estimate, which is the sum of parameter dispersion and error dispersion (Raudenbush & Bryk, 2002). As the error dispersion does not reflect measurement error in HLM, the variance of between-level coefficient estimate is underestimated and so is the standard error. In contrast, the latent variable approach yielded less biased estimates, and

statistical inferences across sampling and the ICC size conditions were more consistent than those of the traditional model, as long as the number of groups is sufficiently large (25 was found to be too small).

The third purpose of this study was to provide empirical illustrations using two subsets of data extracted from PISA (Adams & Wu, 2002). The relation between reading literacy and co-operative learning preference was examined, using a subset of PISA data collected in Korea. A negative, but not statistically significant, cross-level interaction was found between reading literacy and co-operative learning preference. The nonlinear multilevel latent variable model and the traditional HLM approach yielded similar results in that the cross-level interaction estimates were not statistically different from zero in both results.

Unlike the results from the simulation study, the results of empirical applications were not dramatically different in model comparison. One possible explanation is that the predictor variable reading literacy is measured by a large number of well-developed items for these empirical applications, and accordingly, the summed scores are very reliable. However, in other circumstances where less reliable measures (e.g., affective domain measures or teacher instructional variables) are used as predictors or where even a smaller number of people per group are sampled, it is expected to observe more substantial differences between the results from a nonlinear multilevel latent variable model and a traditional HLM. In addition, these two models can also yield divergent statistical inferences even when there are a sufficient size of ICC and a large number of people per group due the substantial elevation of Type I error rates when the traditional HLM is applied. Therefore, a wide range of further empirical applications should be followed.

The improved estimation efficiency, by adopting an MH-RM algorithm for the nonlinear multilevel latent variable models, can contribute to further applications by making the nonlinear multilevel latent variable modeling framework more practical in routine use. However, more varied versions of MH-RM algorithm are worth to be pursued for further research. While the point estimates are easily approximated, the calculating proper standard errors still requires more investigation. For further estimation efficiency with respect to more complex models, two-stage estimation in the framework of MH-RM algorithm is being studied.

REFERENCES

Adams, R., & Wu, M. (2002). *PISA 2000 technical report*. Paris, France: Organization for Economic Cooperation and Development.

Cai, L. (2008). *A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model*. Un-published doctoral dissertation, Department of Psychology, University of North Carolina - Chapel Hill.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis- Hastings Robbins-Monro algorithm. *Psychometrika*, *75*, 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307–335.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700-725.

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*, *32*(3), 252-286.

Goldstein, H., & Browne, W. (2004). Multilevel factor analysis models for continuous and discrete data. In Maydeu-Olivares & M. J. J. (Eds.), *Contemporary Psychometrics* (p. 7270-7274). Mahwah, NJ: Lawrence Erlbarum Associates.

Hastings, W. K. (1970). Monte Carlo simulation methods using Markov Chains and their applications. *Biometrika*, *57*, 97-109.

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). *Multilevel modeling of educational data*. In A. A. O'Connell & McCoach, D. B. (Eds.), (pp. 345–388). Charlotte, NC: Information Age Publishing.

Louis, T. A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society*, *44*(2), 226-233.

Lüdtke, O., Marsh, H., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203-229.

Lüdtke, O., Marsh, H., Robitzsch, A., & Trautwein, U. (2011). A 2 x 2 taxonomy of multilevel latent covariate models: Accuracy and bias trade-offs in full and partial error-correction models. *Psychological Methods*, *16*(4), 444-467.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, *44*, 764-802.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1092.

Muthén, L. K., & Muthén, B. O. (2008). *Mplus 5.0* [Computer software]. Los Angeles, CA.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (Sixth ed.). Los Angeles, CA: Muthén & Muthén.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2012). *nlme: Linear and nonlinear mixed effects models* [Computer software manual]. (R package version 3.1-104).

R Core Team. (2012). R: *A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org (ISBN 3-900051-07-0).

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*, 400-407.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs*, *17*.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.

Call for Manuscripts
KAERA Research Forum Volume 1 Issue 2
"International Comparative Education-Korea and Other"

We are pleased to announce a call for manuscripts for the second issue of the *KAERA Research Forum, "International Comparative Education-Korea and Others."* For this second issue, Dr. Nam-Hwa Kang at Korea National University of Education serves as the Guest Editor and Dr. Soo-yong Byun at Pennsylvania State University as Associate Editor.

KAERA Research Forum gives an opportunity to both established scholars and emerging researchers including graduate students. For established scholars, this is a place where they can share their newest, intriguing ideas still in progress and unfolding. For junior scholars and graduate students, this forum will serve as one of first outlets to share and disseminate their scholarly work. Their initial work presented in this non-refereed online outlet may be revised later and developed into a full manuscript for publication in a refereed journal.

The second issue will includes research studies focusing on "International Comparative Education-Korea and Others." **If you are interested in sharing your research on international comparative studies in relation to Koreans in the world, please submit a complete report via email to the guest editor** (nama.kang@gmail.com) **by February 28th, 2014**. (See below for more detailed information about the research report series and submission process.) If your submission is selected to be included in the issue, you will be notified of the acceptance with further instructions.

Submission Guidelines for Authors Invited to Submit a Manuscript

All submissions must be in Microsoft Word format and conform to the most recent edition of the *Publication Manual of the American Psychological Association* (APA).

Manuscripts should include the following:

1. Title page that includes all authors' names, affiliations, and contact information for the lead author.
2. Body of the paper with the following sections:
    - *Introduction*, which documents the importance of the topic and the purpose of the report.
    - *Appropriate practice for research and/or evaluation*, which uses scholarly literature to describe the method(s) discussed in the paper.
    - *Synopsis of a research study*, which demonstrates the application of appropriate practice in research and/or evaluation.
    - *Implications for research and evaluation*, which includes recommendations for the research and evaluation community when employing the method(s).
3. References

Manuscripts should be 1,500 to 2,500 words, not including the title page and references. All figures and tables should be submitted in APA style.